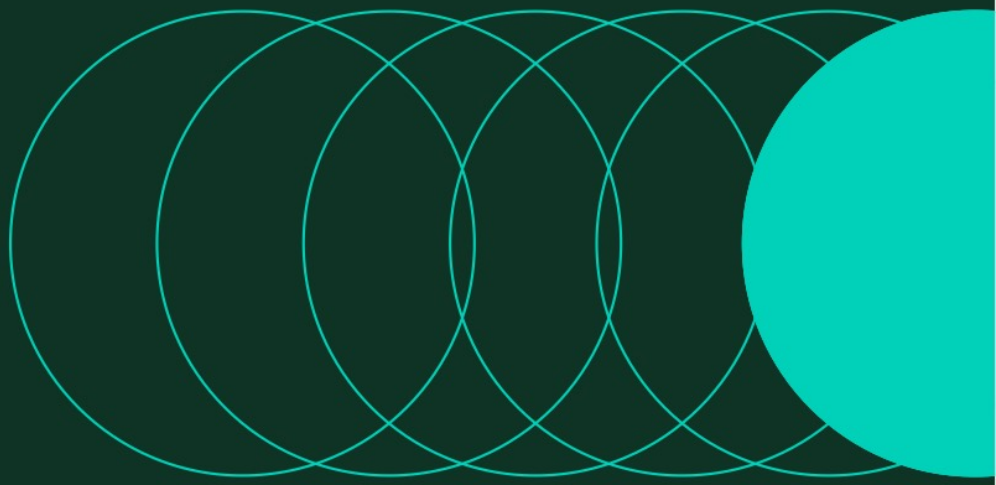
AI Credit Monetization: From Hype to Concrete Blueprint

A Revenue & Finance Executive Guide to Stop Second-Guessing and Avoid Margin Meltdown or Customer Revolt

By Michael Mansard, Principal Director, Subscription Strategy, Zuora
First published in April 2026 – Licensed under CC BY-SA 4.0



ZUORA

Contents

Author's Note:

This article provides a purposefully dense, all-in-one overview for software and AI executives and buyers. Presented as a comprehensive whitepaper rather than a traditional post, it is not intended to be read in one sitting. Please bookmark this page to reference the tools and benchmarks as needed. For those seeking a granular focus, this content is also available broken out into three individual articles centered on the specific tools introduced here.

I would like to thank Steven Forth for his detailed review and contributions, Maciej Kraus and Mélanie Septe for their feedback (among many others!).

EXECUTIVE SUMMARY

Key Takeaways

- **The AI Margin Crisis:** 30% of AI-first software companies have already adopted credit-based models as a monetization mechanism. Credits have clearly become a (trendy) attempt to address the industry margin crisis imposed by the brutal dynamic of AI unit economics.
- **Trendy, But Not New:** Credits aren't an "AI breakthrough." They are a decades-old mechanism borrowed from the Telco and Information-as-a-Service playbooks. Multibillion-dollar firms like Dun & Bradstreet have used credits for years for a key reason: they are an effective tool to bridge the gap between highly volatile costs and unpredictable customer utilization.
- **Monetization Reality, not just Rhetoric:** Native AI leaders like Replit and Higgsfield have already vaulted past the $250M ARR mark using credit-based models. Among incumbents, ServiceNow surpassed $600M in 2025 AI-related ACV and is targeting over $1Bn in 2026, with on-demand credit "Assist Packs" growing at a staggering 50% MoM, accounting for roughly 15% of their total AI revenue.
- **The Triage Tool:** Credits aren't a panacea. Use this article's Credit Model Flash Triage, a simple decision tree to verify if your solution actually has the 4 traits that justify the need to introduce a credit-based model. Or if credits will simply add unnecessary friction and complexity.
- **The 12-Attribute Credit Blueprint:** Most credit models are still built on gut feel, ending up as a random collection of features rather than a deliberate architecture. To end the guesswork, the article introduces a disciplined 12-attribute credit model taxonomy (such as Granularity, Persistency, Portability) to eliminate design blind spots and streamline future iterations.
- **From Trap to Trust:** The market is moving away from rigid, black-box "vendor-first" traps. We put the credit architectures of ServiceNow, Figma, HubSpot, and Higgsfield under the 12-attribute microscope to show how the push toward customer-centric models (e.g. rollovers, extended expirations, and cross-portfolio usage) is steadily raising the stakes for the entire industry.
- **The Hybrid Reality:** Credits do not live in isolation. 60% of AI offers are already hybrid, using a Fixed-Fee Chassis for stability and a Variable Credit Engine for scale. Hybrid is also a powerful lever to bridge models, typically, using seats for "Copilots" (human-led), and platform fees + credits for "Autopilots" (fully autonomous). Hybrid models also come with a significant upside: they yield an 8%–13% growth premium over simpler models.
- **New Ops Playbooks:** Credits aren't a "set-and-forget" pricebook update, but a massive cross-functional lift. This shift requires new muscles across the firm: Sales must navigate complex procurement and lifecycle objections, while Finance must master "revenue breakage" and non-linear revenue recognition. Without a total operational overhaul supported by fit-for-purpose systems, credits will brutally erode customer trust and crush your teams.

"You have no credits left. Do you want to buy more?"

If the memory of a prepaid card running out mid-call, leaving you stranded in a phone booth while trying to reach a relative, triggers a specific kind of panic, you are not alone. By the way, if that sounds like a scene from a period piece, you were likely born after 2000, but fear not: this article is still relevant to you. In the high-stakes world of AI monetization in 2026, this "old-school memory" has indeed resurfaced as the most critical revenue topic in the boardroom.

As AI dismantles traditional SaaS economics and profit margins, the industry has been searching desperately for a stabilizer (if you are interested, more on this in my previous COMPASS articles here). While Outcome-Based Pricing was the undisputed buzzword of 2024, Credits have suddenly emerged as the industry's default, Pavlovian response to the "AI margin crisis".

Adobe made a bold pioneering move, launching "Generative Credits" for Firefly in September 2023. At the time, they were almost alone. But by 2025, the floodgates opened: Salesforce introduced Agentforce's "Flex Credits," ServiceNow and Workday followed suit, and HubSpot launched "HubSpot Credits". More recently, OpenAI began capping usage to drive credit purchases for Sora or Codex, while Microsoft introduced "Agent Mode credits". From start-ups to incumbents, everyone seems to be betting on the credit bucket: as of early 2026, over 30% of AI-native companies have adopted credits as one of their core monetization mechanisms, up from a tiny minority just a year ago.

Yet, this rapid shift raises a critical question: what are credits really about? To move beyond the buzzword, revenue and finance leaders must understand the fundamental "why" behind AI credits, determine when they make sense, how to design them, and how to avoid operational pitfalls. This is the purpose of this article. To do so, we introduce a suite of tools, benchmarks, and case studies organized as follows:

- **Chapter 1:** We begin by defining credits and their core function;
- **Chapter 2:** We trace their lineage back to early heavy hitters like Dun & Bradstreet, who pioneered the normalization of information arbitrage;
- **Chapter 3:** We introduce the Credit Model Flash Triage, a decision tree to determine if your offer is a genuine candidate for credits;
- **Chapter 4:** We walk through the Credit Model 12-Attribute Taxonomy, a framework built to help you design a credit architecture without blind spots. We then stress-test it against the real-world examples of ServiceNow, HubSpot, Figma, and Higgsfield.
- **Chapter 5:** We perform a deep-dive into the 12 attributes through the collective lenses of the CPTO, CRO, and CFO to bridge strategy and operations;
- **Chapter 6:** We provide tactical Credit Migration safeguards and learnings to guide the transition from your current model to a credit-based architecture, including Cursor's cautionary tale.

Note: This is a high-conviction, detailed deep dive. If you are already comfortable with credit fundamentals, skip directly to Chapter 3 (The Credit Model Flash Triage) to assess if your current model is a fit-for-credit, or to Chapter 4 (The Credit Model 12-Attribute Taxonomy) to audit your credit architectural design.

Note: I highly recommend reading the deep dives by my friend and pricing expert Steven Forth. While he wrote many pieces on the topic of credits, his Design Choices in Credit Based Pricing is an essential companion piece for anyone building in this space.

Note: This article is the 4th installment of the COMPASS AI Monetization Framework (repository here). The article introduces an additional set of tools framing and supporting the continuous design decisions of your credit-based model into our collective toolbox.

CHAPTER 1

The latest ‘new’ old idea linked to the AI monetization revolution

While "credits" may seem to some as a modern digital innovation with the current AI hype, they are essentially "old bottles" for a vintage economic wine. Much like the outcome-based models that revolutionized advanced manufacturing decades ago, think Rolls-Royce’s "Power by the Hour" or Michelin’s "Pay-per-Kilometer", software vendors are once again rediscovering another strategy, this time long utilized by telecommunications giants and Data-as-a-Service (DaaS) providers to manage complex resource consumption. Credits are indeed far from new; they are a battle-tested strategy for aligning price with the messy realities of value, cost, and adoption.

Let’s be theoretical nerds for one paragraph (I promise, not more!), before moving to layman’s terms: credits function as a synthetic currency designed to decouple the volatile utility of specific assets from the friction of fiat transactions. In effect, credits are prepaid units of future, unknown entitlement. Think of them as "use-it-however-you-need-it" buckets or "Service IOUs". For the buyer, this secures the right to use "something useful" at an unspecified time, removing the need to predict exactly which services they will consume at the point of purchase. For the seller, credits secure upfront commitment and immediate cash flow, while providing a hedge against the highly variable costs (effectively shielding the business from the margin-eroding risks of unlimited subscription models). **Credit models effectively decouple the act of buying, the act of having, and the act of using.** Finally, by abstracting price from point of consumption, credit models allow both parties to move past the "procurement conundrum".

To bring this theory down to earth and put it into layman’s terms, let’s step back again to the days of the old-school, analog phone booth. Before the era of "unlimited" digital plans, every call carried a wildly different weight. Calling your neighbor cost the provider pennies, while calling a cousin in a foreign country required expensive undersea cables. Crucially, the value of those calls was just as unpredictable as the cost: one day you’re making a high-stakes business deal from a booth in the city, the next you’re just telling your date you’re late for dinner on a remote roadside. Because neither you nor the provider could forecast whether you’d need a three-minute local chat or a thirty-minute international crisis call, a recurring monthly fee would have been either a rip-off for the light user or a bankruptcy risk for the provider.

The solution was the prepaid phone card. By purchasing a card with 100 units, you essentially "froze" the price of communication at that moment; even if the provider raised rates the following week, your card held its original purchasing power. This gave you a portable "inventory of utility" that you could carry in your pocket. The system then weighted your usage: a local call burned one unit, while that international birthday wish burned ten. It was perceived as a fair, transparent way to manage a world where every action had a different cost to the vendor and a different value to you, ensuring you only spent your precious units when the utility of the call justified the price. Credits can thus act both as a value storage currency and a psychological catalyst solving "commitment anxiety," by allowing buyers to engage with complex systems without fearing unpredictable financial exposure.

On a side note, while we have been focused on unit-based credit models, the most elementary model is the Monetary Credit. Think of it as a "deposit-and-draw-down" system directly in prepaid currency. This was typically the model for some prepaid plans in the Telco industry, which functioned as a solvency guardrail to ensure customers were "good for the money" before triggering high-cost usage. Beyond risk management, it can alternatively serve as a commercial incentive: customers trade upfront liquidity for volume discounts, effectively pre-buying an "inventory of utility" at a lower rate. Though it is not a widespread model as of March 2026, Anthropic is a case in point: the company gives users the option to buy "monetary credits" of extra usage that can be used across their products with a volume discount from 10% to 30%. Replit is another: the $18/month subscription grants $20 worth of monthly credits value.

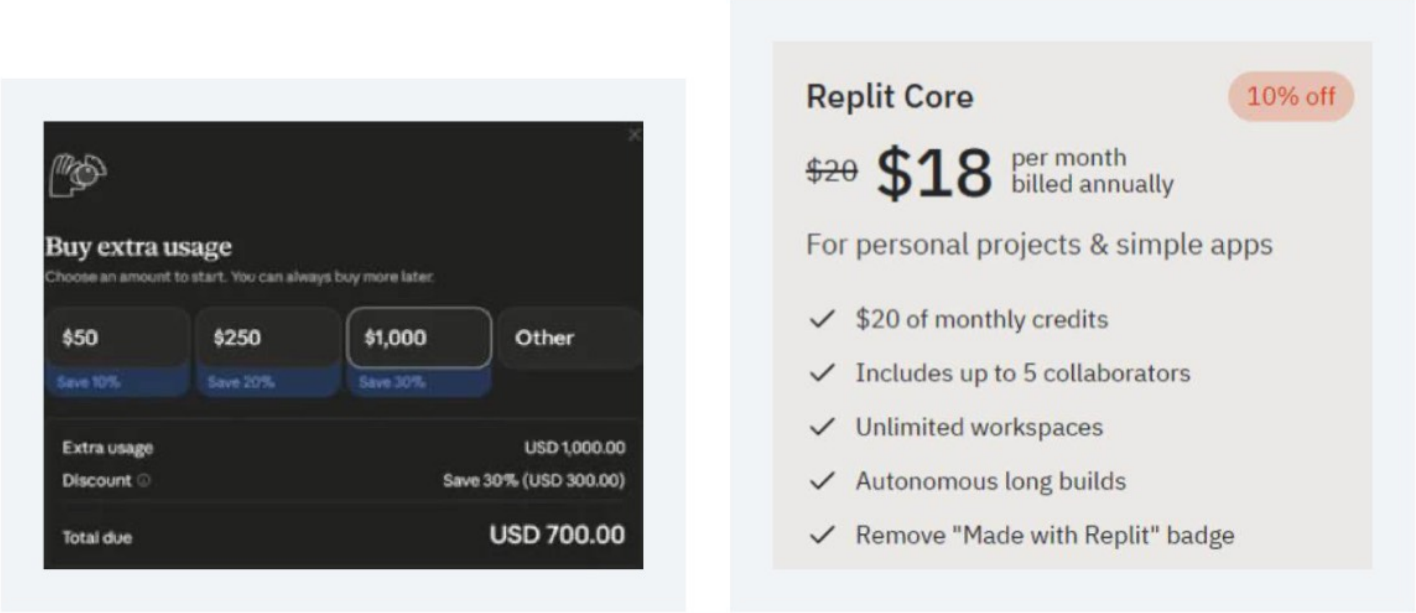


Figure. Anthropic gives users the option to buy "monetary credits" of extra usage. Replit makes it core to its recurring offers.

We could summarize credit-based models with this very simple "loop" flow:

- **Acquisition:** Customer buys credits (via a one-off transaction, a plan, etc.)
- **Allocation:** Credits are stored into the customer credit balance / "wallet" (separating the "Buying" from the "Having")
- **Execution:** User or Agent initiates "service executions"/"work"
- **Evaluation:** Executions are evaluated / rated to quantify credit consumption (e.g., 1 image = 5 credits).
- **Consumption:** Credits are deducted / drawn-down from credit balance / "wallet"

Credit-based models eventually faded in Telco as infrastructure became digital and the marginal cost of a "call" dropped to near zero. SaaS followed a similar pattern: expensive to build but cheap to run. But AI has triggered a reversal of digital economics. In a world of complex AI orchestrations and token-heavy inference, costs are once again highly variable, consumption patterns erratic, and margins volatile. Because the precise value of an inference is often not yet fully understood nor ascertainable by either the buyer or the vendor, the credit model has returned. Not as a relic, but as an essential stabilizer for navigating this new era of economic unpredictability.

Finally, to clear a common misconception: Tokens are not credits. Tokens are the fundamental "atoms" of AI models: the raw, granular measurement of computational activity. When companies charge directly for tokens, they are operating within a "cost-plus" usage-based paradigm. By utilizing credits instead, a company can (1) decouple the act of buying from the moment of use, (2) abstract away the volatile underlying computational cost (the tokens) that consume credits, and (3) realign the invoice with high-level activities and/or outputs and/or outcomes. This also implies that credits are not a pricing metric per se, but a normalized abstraction layer designed to decouple payment from access.

As expert and friend Maciej Kraus aptly frames it:

> “
>
> Tokens = compute unit, credits = commercial abstraction, outcome-based model = value contract.
>
> **Maciej Kraus**
> Managing Partner, Movens Capital

CHAPTER 2

The Pre-AI Credit-Based Models Heavy Hitters

While prepaid credit models faded from the headlines following the decline of legacy telecommunications, they have remained the quiet, multi-billion-dollar engine of the Data-as-a-Service (DaaS) and Information-as-a-Service sectors. For decades, the industry's heavy hitters have utilized credits not as a novelty, but as a primary paradigm for information arbitrage. To decipher and anticipate the trajectory of modern AI 'tokens' or 'credits,' one must first analyze the original architects of this model. Dun & Bradstreet (D&B) stands as a preeminent pioneer and the logical place to start.

Dun & Bradstreet: Snapshot of a Credit Pioneer

Long before "SaaS" entered the corporate lexicon, D&B generated billions in revenue over decades through a sophisticated credit-based system. They primarily sold business risk intelligence and credit reports to finance, procurement, and sales departments. These teams used D&B data to vet the solvency of potential partners or to build targeted lead lists for expansion.

The credit model was born out of the inherent variability of information value. Not all data is created equal, and the "cost to produce" a report also varies wildly based on geography and data accessibility. For example, generating a report on a private company in an emerging market (like China in the early 2000s) required manual investigation and local expertise. Conversely, a standard payment analysis for a US-based firm relied on automated, high-volume data pipelines. A flat dollar price for both would either cannibalize D&B’s margins or overcharge the customer.

Credits solved this by allowing D&B to create a multi-dimensional rate card:

- High-Complexity Report (e.g., China BIR): 50 Units
- Low-Complexity Report (e.g., US Payment Analysis): 8 Units

By pricing in "Units" rather than dollars, D&B effectively decoupled the cost of the data from the cash transaction. While the effective unit price of a credit fluctuated, for instance between $0.80 and $3.00 depending on a customer's total volume, the end-user only perceived the "credit unit cost." This psychological shift was brilliant. It allowed D&B to charge the equivalent of $50 to $150 per report without the friction of a direct dollar-to-data transaction. To the customer, the data felt like a "prepaid utility": already bought, paid for, and ready to be used. The "Unit" also allowed D&B to create a "universal pricing language" for their customers needing international information across 200+ countries.

The other genius of the model lay in its expiration logic used as a counter-intuitive retention lever. Customers typically "procured" a bucket of units (e.g., 10,000 units) through an annual Committed Volume Agreement. If those units weren't used by year-end, they were forfeited. This was a potentially massive margin driver for D&B. However, sales reps used this "breakage" as a powerful reason-to-stay. They would offer to "rollover" unused credits into the next year, but only on the condition of a contract renewal at the same or higher volume. This tactic created a "synthetic floor" for their revenue, contributing to a staggering $500M–$700M in Unearned Revenues on their balance sheets throughout the mid-2000s.

This "per-unit luxury" model eventually faced an existential threat from both competitive disruption and a fundamental shift in how data was consumed. "Digital Attackers" like CreditSafe entered the market with "Unlimited Subscription" models, treating business data as a commodity rather than a metered resource. Simultaneously, the primary use case for D&B data expanded beyond isolated risk assessments into mass Sales and Marketing integrations. When a customer needs to sync, enrich, and clean hundreds of thousands of records within a CRM like Salesforce or Microsoft Dynamics, the legacy "unit-per-report" becomes too unwieldy. To adapt to this high-volume environment, D&B shifted toward record-based, API-driven, and subscription-plus-overage pricing. However, the underlying logic, metering variable-value data consumption, remains conceptually close.

Setting a precedent

While D&B did not invent the idea that information could be metered, it helped normalize the logic that variable-value data should be priced through units, credits, or consumption buckets rather than flat access. That mattered because it gave the market a template: when the cost to produce, retrieve, or enrich information changes by geography, depth, urgency, or risk, the cleanest commercial answer is often not a single subscription price, but a metered rate card.

By the late 1990s and 2000s, this logic was in fact already visible across the major information utilities. Experian, Equifax, and TransUnion built large-scale credit and identity businesses around pulls, reports, and monitored accounts. LexisNexis and other legal-research platforms used transactional units to separate ordinary queries from premium documents or high-cost archives. The pattern was consistent: the customer bought access, but the value was actually consumed one unit at a time, with the highest-margin products reserved for the most specialized, most expensive, or most difficult-to-source information.

It became a precedent for how to monetize asymmetry. A simple domestic firmographic record might be cheap to serve, while a cross-border ownership tree, a private-company intelligence file, or a manually verified risk profile could justify a much higher unit weight. Credits solved the pricing problem without forcing the vendor to disclose the full cost

structure. They also solved the buyer's budgeting problem by turning unpredictable information demand into a pre-committed consumption envelope.

Adapting to the new reality

Fast forward to today, credits did not disappear for the Information-as-a-Service giants. They evolved to coexist with the era of data utility. They still use similar commercial architecture, complemented with modern labels and mechanisms: API calls, records enriched, monitored entities, search bundles, overage packs, platform fees, and enterprise commitments. The largest accounts still negotiate away friction through custom enterprise terms. But at the end of the day, industry leaders still rely on the same basic logic: D&B monetizes report consumption, Experian meters access, and LexisNexis differentiates ordinary use from premium retrieval.

Their credit models, however, were modernized to address new use cases and counter digital attackers like Creditsafe, who disrupted the market with simpler, SaaS-like subscriptions. At D&B, for example, the legacy 'Unit' is now one of many options. It remains a very effective tool to capture value for complex, cross-border, and high-value data, whereas simpler "per report" or capped-unlimited models are used for standardized, high-volume reference data.

Ultimately, this coexistence of models and offers allows former "credit heavy hitters" to satisfy the modern demand for predictable budgeting while retaining the high-margin "information arbitrage" of their original model. What began as a way to price reports became a way to price risk, access, latency, and complexity. That is why the pre-AI credit economy matters: it was the commercial ancestor of today's tokenized inference economy, where the core problem remains the same: how to charge fairly when both the value and the cost of a single interaction is highly variable.

Beyond historical data utilities, giants like ServiceNow and HubSpot, alongside high-growth AI entrants like Cursor, Lovable, and Higgsfield, are now stress-testing these models in real-time. Through constant iteration, they are playing with different levers of credit monetization, the very tenets we will soon dissect: experimenting with rollovers, 'top-up' mechanisms, or shifting consumption windows to name a few.

CHAPTER 3

Navigating Usage Complexity: When Credit Models Make Sense

History shows us that credit has been a response to a margin crisis on multiple occasions. But a response is not a strategy. To address this gap, the COMPASS framework introduces a systematic 12-Attribute Credit Model Taxonomy to decrypt credit's "DNA". But before jumping to solutions, it is important to first understand where credit models truly make sense, and where they fail.

The "Credit Model Flash Triage"

Instead of hype, this article introduces a scientific "Triage Tool" based on four fundamental factors. The goal is to quickly evaluate if your solution is a serious "credit candidate," or if credit-based models are over-engineered for your context and you should instead use more direct monetization models.

To do so, we will use the following formula based on these four factors that should now feel straightforward given our findings so far:

```
CREDIT IDEAL CONTEXT = [COST-TO-SERVE VARIABILITY 🟩] + [CAPABILITY DIVERSITY ⚙️] + [VALUE HETEROGENEITY 💎] + [UTILIZATION UNPREDICTABILITY 🎲]
```

There is no single "correct" answer, but rather a Monetization Spectrum that we address through a weighted triage mechanism.

By applying simple YES / NO responses to each of the four factors, you can map your solution's gravity toward credit.

With the following:

1. **Cost-to-serve Variability (CV):** Is there such variance in your Cost of Goods Sold (e.g., GPU/LLM compute) that it materially impacts your Gross Profit Margin? In the digital realm, we grew accustomed to near-zero marginal cost, but with AI, inference costs are rewriting the math. Obviously, unless a technological breakthrough occurs, this is a strong "YES" for the vast majority of GenAI and Agentic AI offers (Note: if you are interested, I researched the underlying reasons for AI margin impacts for the past 3 years, including the previous COMPASS and GenAI articles that you can find here).
2. **Capability Diversity (CD):** Are the "things" produced or done by your AI solution highly diverse, even for a given customer? You would typically have a "YES" if your solution is providing different AI agents or capabilities that tackle distinct Jobs-to-be-Done (e.g. review accounting errors, prepare the accounting books, chase late payments).
3. **Value Heterogeneity (VH):** Is there a significant ROI gap between your lowest-value and highest-value "things" produced by your solution, even for a given customer? This is a "YES" if your platform delivers a mix of basic utility (e.g., a simple data clean up) and high-impact outcomes (e.g., a strategic lead recommendation).
4. **Utilization Unpredictability (UU):** Are your customers' consumption patterns unpredictable? It is a 'YES' if the majority of your customers are unable to commit to even a very minimal baseline (e.g., an early-stage pilot with unproven volume or innovative use case with no precedent), or if usage is tied to a material level of unpredictable outliers (e.g. a "M&A due diligence agent" for mid-size companies).

And a simple rule of thumb: when in doubt for one of the factors, simply pick "NO".

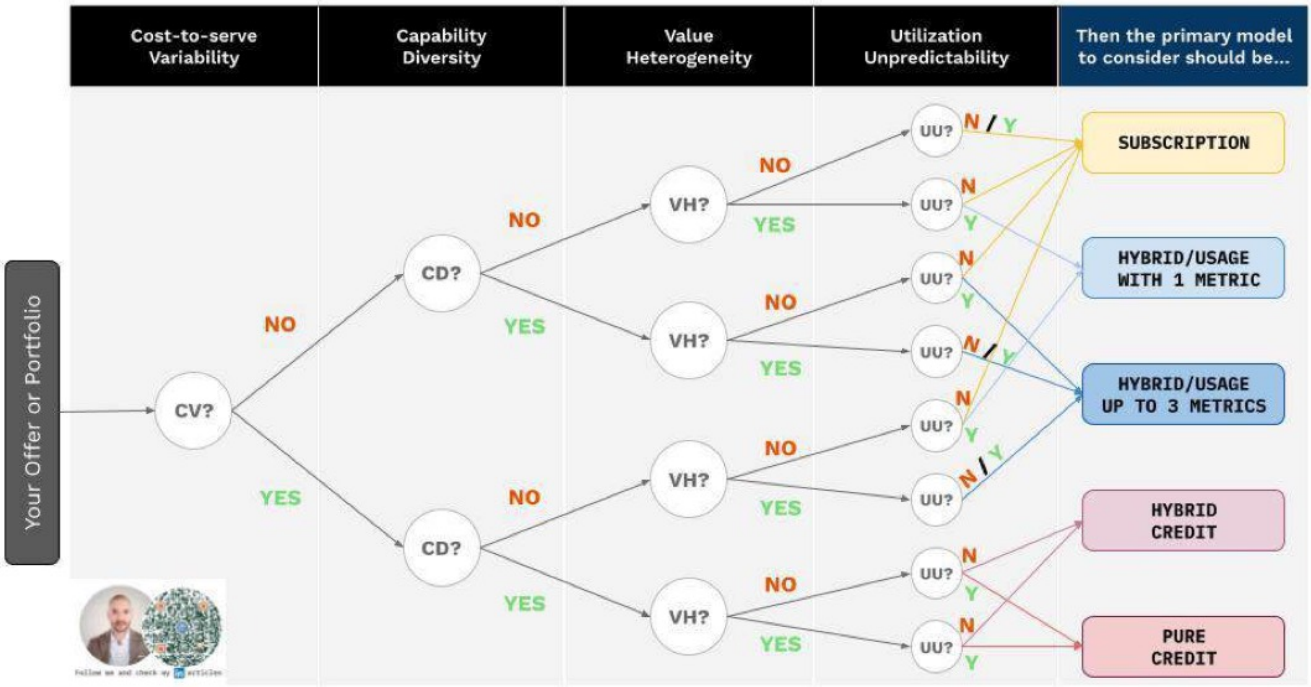


Figure. The Credit Model Flash Triage as a simple decision tree (high-level convenient visual version)

CV Cost-to-serve variability	CD Capability diversity	VH Value Heterogeneity	UU Utilization Unpredictability	Primary Model to consider *Simplified "Decision Tree" vers*	Primary Model to consider *Detailed version*	Nuanced Rationale
✗	✗	✗	✗	Subscription	Simple Subscription	No complexity, no variance to shield; no diversity to bridge.
✗	✗	✗	✓	Subscription	Subscription + pause and/or overage	As usage is lumpy; allows for "pause" or "on-demand" logic to limit churn
✗	✗	✓	✗	Subscription	Tiered Subscription	Value-based or Job-to-be-done tiered packaging / add-ons
✗	✗	✓	✓	Hybrid/Usage with 1 metric	Usage-based with 1 metric	Value-metric; no/limited committed volume due to UU.
✗	✓	✗	✗	Subscription	Simple Subscription	Feature gating, tiered or Job-to-be-done packaging based on which "Agents" are active.
✗	✓	✗	✓	Hybrid/Usage with up to 3 metrics	Usage-based with up 3 metrics (*)	On-demand "Menu" of capabilities/agents to handle spiky usage. Credits as potential fallback.
✗	✓	✓	✗	Hybrid/Usage with up to 3 metrics	Hybrid with up to 3 metrics (*)	Bundled tiers and metric(s) that align with value, with committed volume+overage when possible.
✗	✓	✓	✓	Hybrid/Usage with up to 3 metrics	Usage-based with up to 3 metrics (*)	Advanced value-based pay-as-you-go. If more than 3 metrics, consider synthetic metric or credits. Credit can help avoid frictions if the metrics are uncorrelated(e.g., a buyer needs 100 seats but 0 exports, or 1 seat and 1,000 exports).
✓	✗	✗	✗	Subscription	Tiered Subscription + overage	"Capped" tiered subscription to protect margin, with usage add-on or overage. Usage-based with 1 metric as fallback.
✓	✗	✗	✓	Hybrid/Usage with 1 metric	Usage-based with 1 metric	Pure on-demand unit pricing to hedge margin risk.
✓	✗	✓	✗	Hybrid/Usage with 1 metric	Hybrid pricing with 1 metric	Follows value (e.g., $ per Lead), hybrid committed volume+overage when possible.
✓	✗	✓	✓	Hybrid/Usage with 1 metric	Usage-based with 1 metric	Pure on-demand value-based pricing (compute risk baked in).
✓	✓	✗	✗	Hybrid Credit	Hybrid Committed credits	Credit acts as a "universal protocol"; requires a recurring "floor", hence hybrid.
✓	✓	✗	✓	Pure Credit	On-demand credit	Complex/Volatile/Lumpy; pure elastic consumption and credits as a protocol.
✓	✓	✓	✗	Hybrid Credit	Hybrid Complex Committed credits	Credit acts as a "universal protocol"; rich value-based credit unit mapping. Requires a recurring "floor", hence hybrid.
✓	✓	✓	✓	Pure Credit	On-demand complex credits	Total abstraction is required as a de-risker for the most complex AI offers.

Figure. The Credit Model Flash Triage detailed rationale (table version).

While this tree provides a logical path to a credit model, it should be viewed as a directional compass, not a rigid mandate. As you navigate these branches, keep four principles in mind:

1. **The "Metric-First" Bias:** Most monetization challenges can often be solved by designing packages, bundling, or refining your primary pricing metrics before introducing the friction or complexity of credits. If you can capture value through a tiered subscription, or a handful of clear usage meters, do so. Credits are a solution to complexity; if you can simplify the product offer instead, you win on user experience.
2. **Directional, Not Dogmatic:** This triage tool identifies where the "gravity" of your business model is pulling you. However, the recommended leaf is rarely the only solution. For instance, a "Credit-Led Hybrid" might be the logical choice for a complex platform, but a "High-Value Recurring Fee" might be the better tactical choice for an early-stage startup looking to reduce billing hurdles during initial adoption.
3. **Cost Variability is the Kingmaker:** The Flash Triage makes it visually clear that credit models are not a default; they are a specialized tool. They make the most sense in a limited set of cases where most of the 4 triage criteria overlap. If Cost-to-serve Variability (CV) is absent, credits are almost exclusively an elective choice for UX abstraction that must be carefully weighted. If CV is present and paired with diversity or unpredictability, credits transition from a "nice-to-have" to a "must-have" margin shield.
4. **A Journey, Not a Destination:** Pricing is dynamic. As your AI capabilities mature and your costs stabilize, the decision tree actually implies a "reverse path." Stabilization creates the opportunity to migrate away from credits and back toward simpler, more direct monetization models that offer even less friction to the customer.

KEY TAKEAWAY

Use this tool to ensure credits are a fit in your specific case. If they don't solve more than one of the four core triage questions, they risk becoming an "unnecessary abstraction" adding operational, financial or UX burden.

Applying the Credit Model Flash Triage with a concrete example

To ground this triage, let's consider a hypothetical case study: OmniSMB, a foundational System of Record for the mid-market. Their clients are lean finance teams operating in high-compliance environments where operational leverage through AI is the only viable path to scale. The company is currently activating a strategic "Agentic Layer" across its legacy footprint; while the initial rollout targets core accounting, the architecture is designed to eventually automate the entire finance, tax, and HR lifecycle.

For OmniSMB's specific platform, we evaluate four key factors to determine the optimal monetization architecture:

1. **[Cost-to-serve Variability= YES]:** The suite utilizes high-compute LLMs for deep ledger auditing and fraud detection. Because these inference costs fluctuate based on data volume, analysis depth, and exceptions encountered, the vendor must protect its Gross Margin.
2. **[Capability Diversity = YES]:** The platform provides distinct agents for Accounts Payable, Cash Flow Forecasting, and Tax Compliance. These are three different Jobs-to-be-Done and utilities under one roof.
3. **[Value Heterogeneity = NO]:** In this specific SMB context, the perceived ROI is relatively uniform. Due to the mission-critical nature of financial data, the CFO maintains 'human-in-the-loop' oversight for every decision. This continuous involvement creates a shared attribution of value, positioning the AI as a reliable productivity catalyst across all tasks rather than a source of high-variance ROI. As a result, the customer views each agent primarily as a critical "productivity booster" that automates a significant chunk of the finance team's work, rather than one agent delivering 10x the value of another
4. **[Utilization Unpredictability = YES]:** Small businesses don't audit their books daily; usage is bursty. It follows predictable events like tax season or closing, as well as unpredictable outliers triggered by identified anomalies.

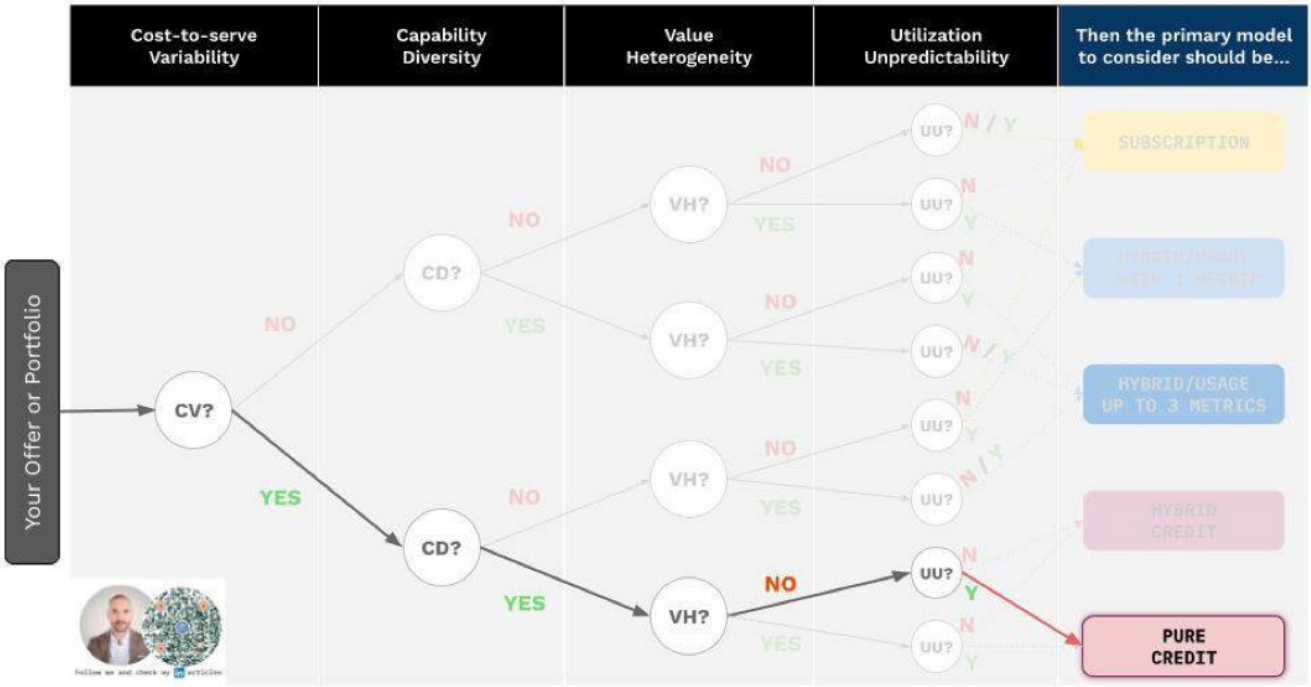


Figure. Applying the Credit Model Flash Triage to the OmniSMB example by following the decision tree until the final node.

The Flash Triage validates OmniSMB as a prime candidate for a credit model. On a side note, considering their specific context, seeding a minimal baseline of credits linked to their customers' user count and/or platform could be relevant to encourage feature exploration. We unpack the specific mechanics of how to calibrate this balance in the next section of this chapter called "Rules of Thumb: Blending and Rightsizing your Credit Strategy".

A well-designed credit-based monetization model can therefore help achieve the following:

1. "Margin Shield" ensures super-users are charged appropriately.
2. "Universal Protocol" allows users to move between features without managing dozens of different meters, while enabling you to rapidly monetize new "things".
3. "Value Tracker" acts as a proxy to capture value surplus without overcharging for basic utility.
4. "Smart Decoupler" separates the pain of payment from the point of usage, eliminating transaction fatigue that can stifle engagement.

Shades of Gray: Why Hybrid Is Gaining Traction

Success in the Agentic Shift is increasingly found in Hybridization. Our Flash Triage is intentionally built on a spectrum of options to mirror this reality: monetization is rarely a binary choice between "Subscription" and "Credit", but rather a landscape of shades of gray. This isn't merely a transactional draw-down of credits; it is a layered architecture.

We are seeing best-in-class organizations move away from "all-or-nothing" models to orchestrate consumption metrics alongside per-seat or platform-based subscriptions, capturing value at every stage of the AI-human workflow. In fact, **close to 60% of AI credit offers are already hybrid.**

ServiceNow provides an interesting illustration of this layering concept. To procure ServiceNow's Now Assist Credits:

- Customers typically upgrade to a "Plus" tier license, which bundles a specific credit entitlement based on their total seat count.
- Alternatively, customers who prefer to keep their current contract can directly purchase Now Assist Add-on ("Assist Packs") to layer a pool of credits onto an existing Pro or Enterprise environment.
- This monetization strategy propelled Now Assist to over $600 million in ACV by year-end 2025, with a clear trajectory toward a $1 billion target in 2026 driven by a 60% pricing uplift over non-AI tiers.
- While seat-based tiers provide the predictable revenue floor, credit-based "Assist Packs" are the fastest-growing component, expanding at 50% month-over-month as customers shift toward high-frequency agentic workflows and currently represent an estimated 10-15% of total AI-related revenues.

HubSpot is another eye-opening, modern blueprint showing what hybridization looks like:

- Tiered Entitlements: Customers receive a baseline credit allowance tied to their highest subscription tier (e.g., 10,000 per month for Enterprise).
- Top-up Elasticity: If baseline limits are reached, AI credit add-ons and "top-up" packs act as the safety valve.

Note: Outcome-Based Evolution: As of March 2026, HubSpot has leapfrogged simple credits for high-value autonomous functions. By charging $0.50 (in credit equivalent) per resolved conversation for the Customer Agent and $1 (in credit equivalent) per recommended lead for the Prospecting Agent, they have effectively aligned the invoice with the customer's specific ROI moment.

Finally, Higgsfield, an AI-native professional "director-level" video generation platform is building its revenue architecture from the ground up without a seat-based legacy:

- The company enables its customers to access credits via five distinct paid subscription tiers (Basic, Pro, Ultra, Business, Enterprise).
- While the Basic and Pro tiers offer fixed monthly allowances (200 and 1,000 credits respectively), the Ultra, Business, and Enterprise tiers have configurable recurring allowances, providing the enterprise-grade flexibility required for high-volume production.

Note: In Chapter 4 of this article, you can find a forensic assessment of ServiceNow, HubSpot, Higgsfield, as well as Figma through the 12-attribute taxonomy of credit-based lens.

The hybrid approach offers customers predictable spending through a base subscription while allowing them to leverage the dynamic capabilities of Agentic AI. While my friend and expert Steven Forth calls it a "layered cake", ServiceNow's CEO Bill McDermott describes its own hybrid model as a 'Goldilocks scenario'.

Beyond customer-centricity, these hybrid approaches are useful to help close the revenue volatility gap. They capture the upside of high-volume usage without sacrificing the recurring revenue predictability that investors demand. One last upside: hybrid models have a proven track record in delivering superior economics, in the range of 8%-13% faster growth through both better acquisition and expansion (back in 2022, in an article called "The Future of GenAI Pricing Metrics and Models", I shed light on key benchmarks that reveal the strong economics of hybrid monetization models. You can access it in this repository if you are curious to learn more.)

Rules of Thumb: Blending and Rightsizing your Credit Strategy

If Hybridization is the strategic target, the immediate hurdle is the actual calibration of the mix. To move from a conceptual "layered architecture" to a functional market entry, leaders can use the following heuristics as foundational guideposts. While by no means an exhaustive set of rules, they offer a practical starting point for your initial calibration.

Metric Blending "Rule of thumb"

The decision to blend credits with seats or a platform fee typically mirrors the level of agency your AI provides:

- The "Co-Pilot" Blend (Credit Allowance + Seats): Opt for this when End Users still provide the primary value in a given workflow, i.e. the classic "human-in-the-loop" scenario (e.g., an accountant using AI to accelerate a month-end close). Seats capture the value of the human collaborator, while credits meter the AI's specific utility.
- The "Agentic" Blend (Credit Allowance + Platform or Direct Subscription): Shift to this model when an AI agent is designed to mechanically reduce the number of seats in the long run and/or is fundamentally reinventing the workflow. In this scenario, seats become a legacy metric; you are instead monetizing the platform's autonomous capability.

Rightsizing Credit Allowance "rule of thumb"

Setting the initial "bucket size" is a balancing act between encouraging adoption and avoiding the "zombie liabilities" of unused breakage. To navigate this, we use the Pareto Principle as a baseline heuristic for calibration, allowing you to rightsize allotments based on your confidence in the forecast.

- Minimal Allotment (The Conservative Floor): When usage forecasting and commitment visibility are limited, the risk of under-consumption (and thus customer frustration) is high. A typical rule of thumb is to lean toward pure pay-as-you-go or a floor commitment set at **20% of your conservative forecast consensus.**
- Baseline Allotment (The Operational Standard): When usage forecasting is possible and commitment is feasible. A typical rule of thumb is to **target an 80% floor commitment** of the forecast. This provides enough "skin in the game" to drive enterprise adoption while maintaining a 20% buffer to handle the "spiky" consumption patterns inherent in AI. Such calibration prevents the friction of "over-storing" unused credits which is a recipe for churn, as well as the margin erosion of over-discounting, while providing your teams a strategic window to reconnect and expand the partnership before the customer ever hits an overage.

Beyond the Credit Model Flash Triage, make use of the full COMPASS framework

If the Flash Triage signals a fit for credit, you can use my extensive COMPASS Framework toolset to build your formal architecture:

1. Identify blind spots using the 12-attribute design model (introduced next).
2. Validate your specific credit metrics via the 3x3 Matrix.
3. Stress-test the revenue model with the "7-Fold Star" to ensure balance between vendor margins and buyer value.

CHAPTER 4

The 12-Attribute Model of Credit Model Architecture

Why is a model needed?

Now that we better understand what credits really are, why they are making a comeback, and when they make sense, it is important to go one level down. As with packaging, there isn't a unique flavor or type of credit: credits come in many shapes or forms. For example, some credits need to be entirely consumed during a given period, while others allow roll-over to the next period. Or some credits are automatically topped up on a monthly basis while others require manual top-ups. And these are only very basic examples.

In fact, after assessing 100+ AI offers these past years, my conviction is that we can boil credit design down to 12 attributes organized in 4 pillars. These 12 attributes are the building blocks of any credit model, and this framework is an attempt to build their taxonomy, as I did back in early 2024 for the "Pricing Metrics Spectrum" that is now used by many companies.

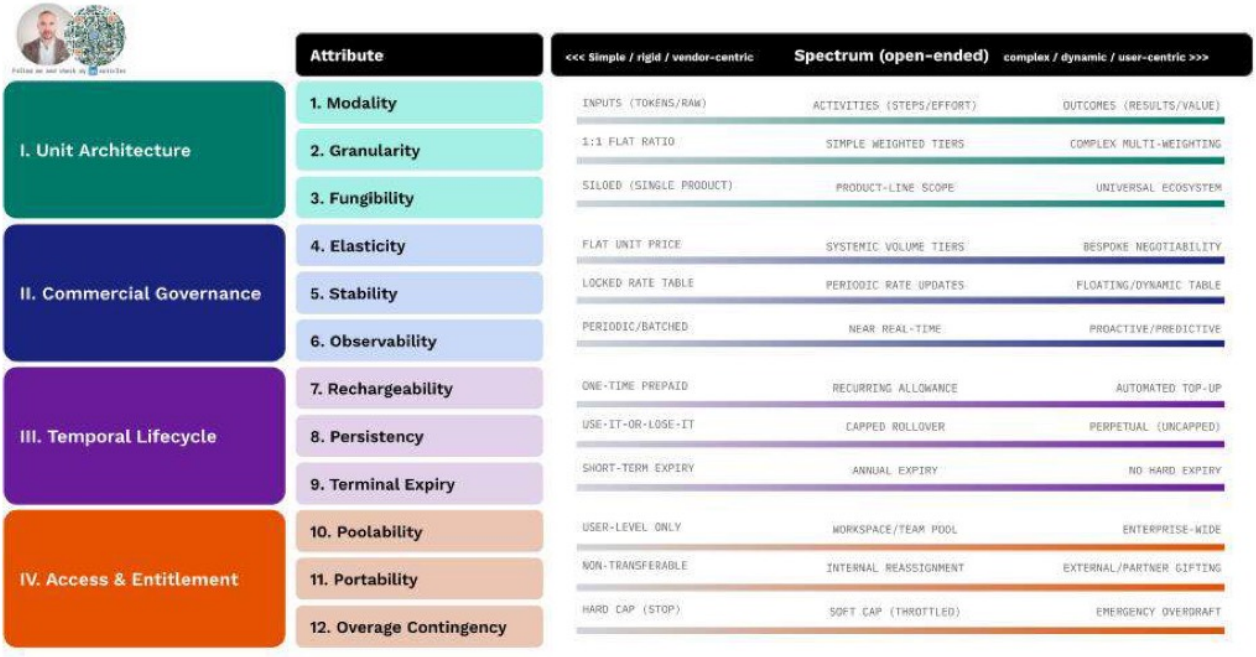


Figure. COMPASS' 12-Attribute Model of Credit Model Architecture and its associated spectrum (open-ended).

Let's put this framework to the test by concretely illustrating with 4 visible AI offers on the credit-based model design decisions they have made as of February 2026. In light of these examples, we will then dive deeper into the real-life implications for executives, especially Product, Revenue and Finance, of each of the framework's attributes design decisions.

Putting the model into action with four real-life AI credit offers (as of March 2026)

To ground once again theory into practice, let's examine how four eye-opening AI credit models are currently configured: ServiceNow, Figma, HubSpot and Higgsfield. These snapshots represent the state of play as of March 2026. Consider these high-resolution illustrative snapshots of a moving target, synthesized from the best public intel available in a market where the market changes by the week.

Case #1: ServiceNow "Now Assist" credits (as of March 2026)

ServiceNow has transitioned from a help-desk utility to an autonomous operating system for the modern enterprise. Its AI agents move beyond simple assistance to execute end-to-end workflows across IT, HR, and Customer Service, fundamentally shifting the platform from a system of record to a system of action.

Note: This framework focuses exclusively on the structural architecture of the credit-based revenue model. Consequently, it purposely excludes customer acquisition and growth tactics, such as Freemium/Trials (initial grant mechanics) or Virality/Referral Incentives (earned credit logic), as standalone attributes. These motions are however indirectly covered through two of the framework's existing Rechargeability and Liquidity attributes. You can read two interesting articles on this : (1) Steven Forth's Design Choices in Credit Based Pricing and (2) a previous COMPASS Framework research on how Agentic AI is changing free trial strategies.

Note: in the same spirit of focus, the framework purposely excludes traditional product or services contractual attributes such as contract terms or payment terms. Those, however, remain critical attributes to define when designing an offer.

Note: As a reminder, ServiceNow's Now Assist surpassed $600M in 2025 ACV and is scaling toward a $1B 2026 target, with consumption-based "Assist Packs" emerging as the fastest-growing lever at 50% MoM to comprise an estimated 10-15% of total AI revenue.

servicenow

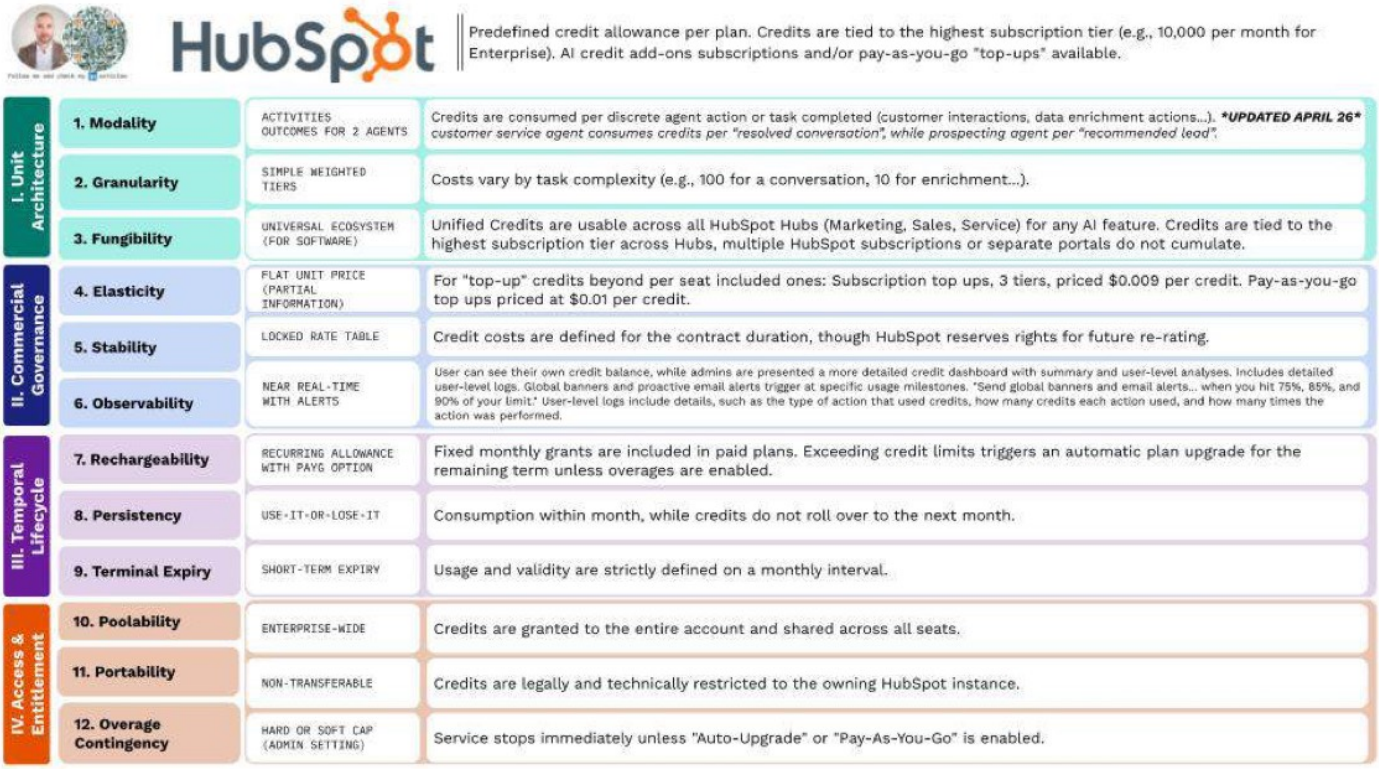
To procure ServiceNow Assist Credits, customers typically upgrade to a "Plus" tier license, which bundles a specific credit entitlement based on their total seat count. If customers prefer to keep their current contract, they can purchase Now Assist Add-on SKUs to layer a pool of credits onto an existing Pro or Enterprise environment. Finally, for those just testing the AI waters, the AI Starter Pack provides a fixed, entry-level bundle of credits and seats designed specifically for pilot programs.

I. Unit Architecture	1. Modality	ACTIVITIES	Cost is calculated based on "actions" performed within a specific skill.
	2. Granularity	COMPLEX WEIGHTED TIERS	Uses a variable "Assist ratio": Simple skills (e.g., summary) can cost 1 assist, complex workflows can cost 25+. Relatively rich/granular list of "Assist ratios" (60+ rate cards). Advanced agentic and custom skills are metered at runtime based on reasoning steps and logic depth.
	3. Fungibility	UNIVERSAL ECOSYSTEM (FOR SOFTWARE)	AI Credits are now pooled across all ServiceNow product families (ITSM, CSM, HRSD, etc.). Software only, not Pro Serv as of today.
II. Commercial Governance	4. Elasticity	BESPOKE (FOR ENTERPRISE) VOLUME TIERS (FOR OTHERS)	Customers purchase specific "entitlement packs" defined by their contract, negotiated with Account Executive. The "Assist Ratios" are tied to the Service Description active at the time of your signing.
	5. Stability	LOCKED RATE TABLE (PARTIAL INFO)	Credit costs are defined for the contract duration.
	6. Observability	PERIODIC/BATCHED	Detailed dashboard that helps understand feature adoption, skill performance, and user behavior. Dashboard data are however "updated nightly". No proactive alerts/thresholds yet, but stated in roadmap.
III. Temporal Lifecycle	7. Rechargeability	ONE-TIME PREPAID	Ability to buy additional credits / negotiate via Sales Rep (i.e. no automatic top-ups nor switch to pay-as-you-go).
	8. Persistency	USE-IT-OR-LOSE-IT	Unused assists do not roll over past the defined subscription term.
	9. Terminal Expiry	ANNUAL EXPIRY	Compliance and usage are typically monitored over a rolling 365-day "burn" contract window.
IV. Access & Entitlement	10. Poolability	ENTERPRISE-WIDE	Consumption is monitored at an "account-level summary" across all instances (Prod/Non-Prod). Any licensed user on the instance can technically trigger an AI skill and draw from that same credit bucket
	11. Portability	INTERNAL REASSIGNMENT EXTERNAL/GIFTING (MSPS)	Usage is fungible across any instance connected to the organizational account. ServiceNow allows MSPs to manage a "Universal Pool" of Assists across multiple of their customer instances. In some 2026 contracts, the MSP might be able to "pass through" a starter bucket of their own Assists to a new client. It is aimed at proving the ROI of a specific AI skill (like Summarization) before the client commits to their own Assist Pack.
	12. Overage Contingency	SOFT CAP (THROTTLED)	No hard cap to ensure service continuity. System monitors compliance; exceeding limits requires renegociation with sales rep. Customer admin can also use observability dashboards to make the decision to deactivate certain skills manually to slow the burn.The AI Control Tower enforces governance through departmental "circuit breakers" that halt execution once spend ceilings are reached. Complementing this, Now Assist Guardian utilizes rate-limiting to provide a technical safety net against "runaway agents" and rapid credit depletion.

Figure. COMPASS 12-Attribute Model of Credit Model Architecture — ServiceNow Assist credits

Case #2: HubSpot (as of March 2026)

With the "Breeze" AI suite, HubSpot is reimagining the CRM as an active participant in the growth cycle. By deploying specialized agents for prospecting, content remixing, and data enrichment, it allows lean teams to execute sophisticated, multi-channel marketing strategies with industrial-scale efficiency.



HubSpot

Predefined credit allowance per plan. Credits are tied to the highest subscription tier (e.g., 10,000 per month for Enterprise). AI credit add-ons subscriptions and/or pay-as-you-go "top-ups" available.

I. Unit Architecture	1. Modality	ACTIVITIES OUTCOMES FOR 2 AGENTS	Credits are consumed per discrete agent action or task completed (customer interactions, data enrichment actions...). ***UPDATED APRIL 26**** *customer service agent consumes credits per "resolved conversation", while prospecting agent per "recommended lead".*
	2. Granularity	SIMPLE WEIGHTED TIERS	Costs vary by task complexity (e.g., 100 for a conversation, 10 for enrichment...).
	3. Fungibility	UNIVERSAL ECOSYSTEM (FOR SOFTWARE)	Unified Credits are usable across all HubSpot Hubs (Marketing, Sales, Service) for any AI feature. Credits are tied to the highest subscription tier across Hubs, multiple HubSpot subscriptions or separate portals do not cumulate.
II. Commercial Governance	4. Elasticity	FLAT UNIT PRICE (PARTIAL INFORMATION)	For "top-up" credits beyond per seat included ones: Subscription top ups, 3 tiers, priced $0.009 per credit. Pay-as-you-go top ups priced at $0.01 per credit.
	5. Stability	LOCKED RATE TABLE	Credit costs are defined for the contract duration, though HubSpot reserves rights for future re-rating.
	6. Observability	NEAR REAL-TIME WITH ALERTS	User can see their own credit balance, while admins are presented a more detailed credit dashboard with summary and user-level analyses. Includes detailed user-level logs. Global banners and proactive email alerts trigger at specific usage milestones. "Send global banners and email alerts... when you hit 75%, 85%, and 90% of your limit." User-level logs include details, such as the type of action that used credits, how many credits each action used, and how many times the action was performed.
III. Temporal Lifecycle	7. Rechargeability	RECURRING ALLOWANCE WITH PAYG OPTION	Fixed monthly grants are included in paid plans. Exceeding credit limits triggers an automatic plan upgrade for the remaining term unless overages are enabled.
	8. Persistency	USE-IT-OR-LOSE-IT	Consumption within month, while credits do not roll over to the next month.
	9. Terminal Expiry	SHORT-TERM EXPIRY	Usage and validity are strictly defined on a monthly interval.
IV. Access & Entitlement	10. Poolability	ENTERPRISE-WIDE	Credits are granted to the entire account and shared across all seats.
	11. Portability	NON-TRANSFERABLE	Credits are legally and technically restricted to the owning HubSpot instance.
	12. Overage Contingency	HARD OR SOFT CAP (ADMIN SETTING)	Service stops immediately unless "Auto-Upgrade" or "Pay-As-You-Go" is enabled.

Figure. COMPASS 12-Attribute Model of Credit Model Architecture — HubSpot credits

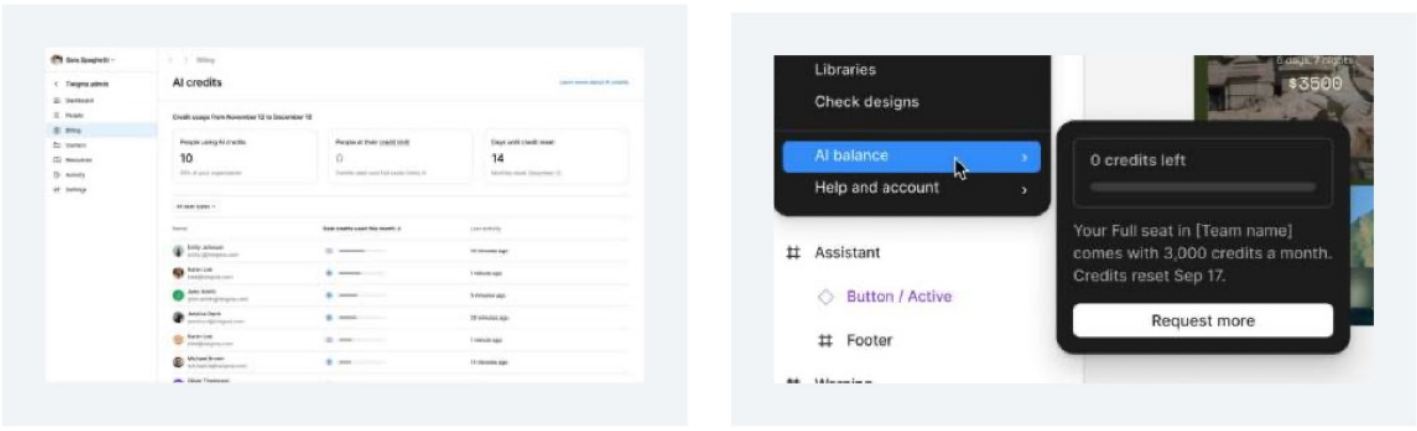


Figure. Examples of how HubSpot provides credit reporting back to admin (left) and users (right).

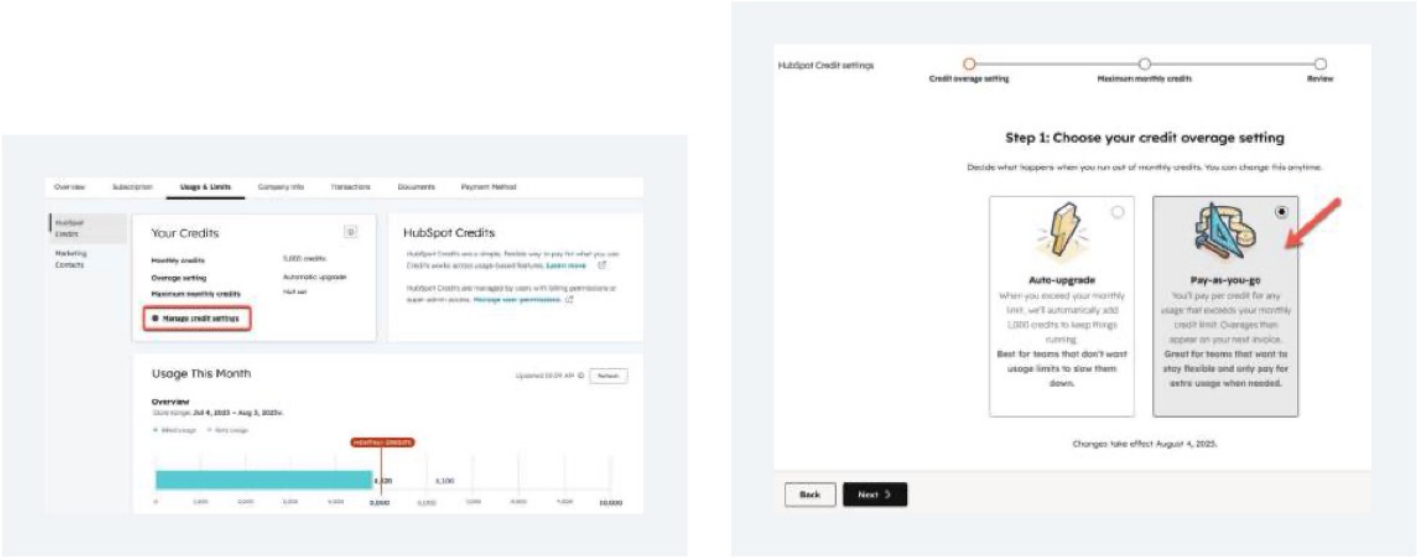
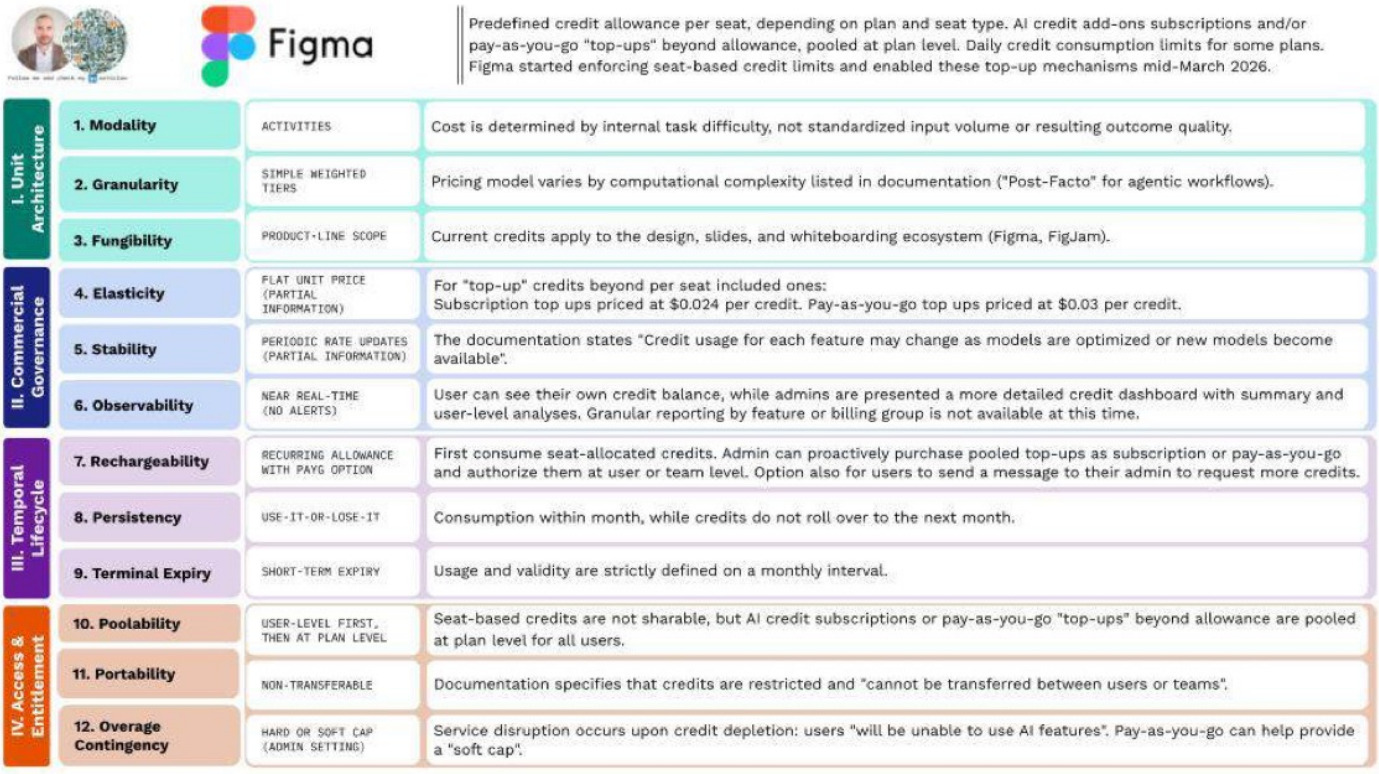


Figure. Examples of how HubSpot empowers admins to decide their credit overage management strategy.

Case #3: Figma (as of March 2026)

Figma is evolving from a collaborative canvas into a generative engine that reduces the distance between design intent and functional code. Through "Figma Make," teams can instantly transform natural language prompts into high-fidelity, interactive prototypes that behave like live applications. By adopting the Model Context Protocol (MCP), Figma is effectively turning the design system into a machine-readable "source of truth," allowing external AI agents to navigate and build directly from the canvas.



Figma

Predefined credit allowance per seat, depending on plan and seat type. AI credit add-ons subscriptions and/or pay-as-you-go "top-ups" beyond allowance, pooled at plan level. Daily credit consumption limits for some plans. Figma started enforcing seat-based credit limits and enabled these top-up mechanisms mid-March 2026.

I. Unit Architecture	1. Modality	ACTIVITIES	Cost is determined by internal task difficulty, not standardized input volume or resulting outcome quality.
	2. Granularity	SIMPLE WEIGHTED TIERS	Pricing model varies by computational complexity listed in documentation ("Post-Facto" for agentic workflows).
	3. Fungibility	PRODUCT-LINE SCOPE	Current credits apply to the design, slides, and whiteboarding ecosystem (Figma, FigJam).
II. Commercial Governance	4. Elasticity	FLAT UNIT PRICE (PARTIAL INFORMATION)	For "top-up" credits beyond per seat included ones: Subscription top ups priced at $0.024 per credit. Pay-as-you-go top ups priced at $0.03 per credit.
	5. Stability	PERIODIC RATE UPDATES (PARTIAL INFORMATION)	The documentation states "Credit usage for each feature may change as models are optimized or new models become available".
	6. Observability	NEAR REAL-TIME (NO ALERTS)	User can see their own credit balance, while admins are presented a more detailed credit dashboard with summary and user-level analyses. Granular reporting by feature or billing group is not available at this time.
III. Temporal Lifecycle	7. Rechargeability	RECURRING ALLOWANCE WITH PAYG OPTION	First consume seat-allocated credits. Admin can proactively purchase pooled top-ups as subscription or pay-as-you-go and authorize them at user or team level. Option also for users to send a message to their admin to request more credits.
	8. Persistency	USE-IT-OR-LOSE-IT	Consumption within month, while credits do not roll over to the next month.
	9. Terminal Expiry	SHORT-TERM EXPIRY	Usage and validity are strictly defined on a monthly interval.
IV. Access & Entitlement	10. Poolability	USER-LEVEL FIRST, THEN AT PLAN LEVEL	Seat-based credits are not sharable, but AI credit subscriptions or pay-as-you-go "top-ups" beyond allowance are pooled at plan level for all users.
	11. Portability	NON-TRANSFERABLE	Documentation specifies that credits are restricted and "cannot be transferred between users or teams".
	12. Overage Contingency	HARD OR SOFT CAP (ADMIN SETTING)	Service disruption occurs upon credit depletion: users "will be unable to use AI features". Pay-as-you-go can help provide a "soft cap".

Figure. COMPASS 12-Attribute Model of Credit Model Architecture — Figma credits

Figure. Examples of how Figma explains its credit-model allotment and charging models to customers.

Case #4: Higgsfield (as of March 2026)

Higgsfield brings professional "director-level" control to the world of AI-generated video. By prioritizing camera physics, character consistency, and multi-model flexibility, it allows brands to move past generic clips into high-fidelity, orchestrated visual storytelling.

Higgsfield

Five distinct paid tiers (Basic, Pro, Ultra, Business, Enterprise).
Basic and Pro have fixed credit allowances of monthly credits (200 and 1000 respectively).
Ultra, Business and Enterprise have configurable allowances within ranges.

I. Unit Architecture	1. Modality	ACTIVITIES	Cost is determined by compute effort (duration, model, resolution, complexity...).
	2. Granularity	SIMPLE WEIGHTED TIERS	Significant multipliers apply: 1080p costs ~1.5x base; 4K costs ~3x. Advanced models cost 40–70+ credits vs 15–25 for basic.
	3. Fungibility	PRODUCT-LINE SCOPE	Credits are universal across the Higgsfield media suite (Video, Image, "Soul" characters, and Lipsync).
II. Commercial Governance	4. Elasticity	SYSTEMIC VOLUME TIERS	Degressive tiered credit price (more volume = lower price per credit).
	5. Stability	FLOATING/DYNAMIC TABLE	"Company may change Credit terms at any time, and the value of Services obtainable with Credits is subject to change at Company's sole discretion".
	6. Observability	NEAR REAL-TIME	Users access credit balance via the account dashboard.
III. Temporal Lifecycle	7. Rechargeability	RECURRING ALLOWANCE	Credits are refreshed monthly for paid tiers. Users can change plan at any time to readjust their credit consumption. Users can also buy one-off "credit packs".
	8. Persistency	USE-IT-OR-LOSE-IT	Unused subscription credits do not roll over to the next billing cycle and expire at the end of each month.
	9. Terminal Expiry	SHORT-TERM EXPIRY	Usage and validity are strictly defined on a monthly interval.
IV. Access & Entitlement	10. Poolability	WORKSPACE/ TEAM POOL	Shared credit pool by default for multi-users plans (Business, Enterprise).
	11. Portability	NON-TRANSFERABLE	Credits are strictly siloed. "Credits have no cash value, are non-transferable, and non-reloadable."
	12. Overage Contingency	HARD CAP	Consumption leads to immediate service pause until manual pack purchase or reset.

Figure. COMPASS 12-Attribute Model of Credit Model Architecture — Higgsfield credits

High-level summary of the four real-life examples

We analyzed four distinct real-life examples that demonstrate the strategic versatility of credit-based architecture. While the four companies are mapped against our 12-attribute spectrum in the visualization below, we can highlight some findings:

- ServiceNow and Figma serve as important examples of the "Legacy Bridge" model, using credits to link traditional seat-based stability with high-marginal-cost AI agents.
- HubSpot has emerged as an eye-opening illustration for "Hybrid Blending," seamlessly orchestrating credits within a tiered platform ecosystem.
- It is also worth noting that ServiceNow has by far the most detailed Granularity, mapping over 60+ distinct rate cards, while at the same time, it remains in a nascent stage regarding real-time Observability.
- Finally, we analyze Higgsfield as the "AI-Native Foundation." Here, credits are not an elective add-on but a foundational synthetic currency required to manage the massive compute volatility inherent in video generation.

What's more, the velocity of iteration within these models is unprecedented compared to SaaS. In a matter of months:

- ServiceNow transitioned toward a 365-day "burn" model, while also significantly increasing Fungibility across its product suites.
- HubSpot leapfrogged simple activity/output-based credits to introduce Outcome-Based pricing for its specialized agents, charging specifically for "resolved conversations" or "recommended lead". This hybrid approach allows HubSpot to mix granular activity-based pricing for utility tasks with high-value, result-oriented pricing for autonomous functions.

But the real "silver lining" of these case studies is the direction of these changes. We are witnessing a clear shift from rigid, vendor-centric "black box" models toward transparent, customer-centric architectures. By leaning into attributes like Capped Rollovers, Extended Persistency and Universal Fungibility, companies are removing buying frictions while fostering utilization. This direction offers a powerful path to move from cost-plus towards value-based monetization. In the agentic era, increased flexibility and transparency are no longer just "nice-to-haves"; they are intrinsic to the value proposition itself.

These are just four examples

As mentioned in previous articles, AI has introduced more change in monetization in 3 years than in 50 years. The AI-credit trend is no exception. For example, in 2026, ServiceNow moved from ad-hoc AI-credit allowances to a 365-day "burn" model where credits expire after one year, while Lovable pivoted away from forced subscription upgrades to a 'top-up' model where users can purchase on-demand credit packs.

To conclude this chapter, let's call out three important points.

- First, there is no right or wrong answer regarding these 12 attributes: they are context dependent. Your context includes information ranging from the customer segments and buyer persona you're serving, all the way through your current assessment of the value you can deliver. As a result, the design decisions you're making today can probably change in 6 months based on new insights. The "wrong decision" would be to blindly ignore these attributes or discover them too late at your own expense.

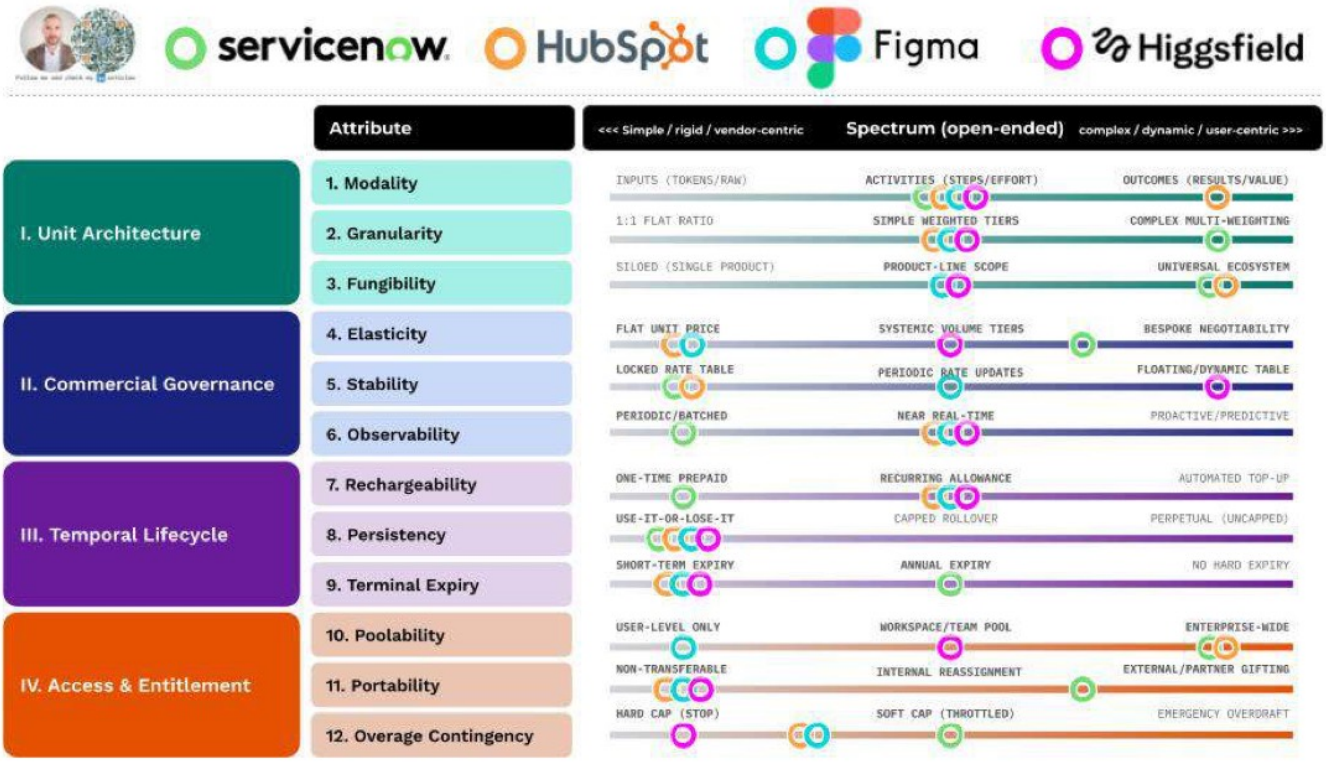


Figure.

- Second, these 12 attributes should in fact be treated as "levers" that need to align with the customer experience you want to design, carefully balancing benefits and risks between you and your customers. If your customers need a PhD to decipher your table, that is clearly not a desirable situation. If your customers pre-pay a lot of credits, but have 80% unused at the end of the period and lose them, even though you made a lot of money, you can expect them to churn soon. Similarly, if your customers can indefinitely store credits while you have rising costs of infrastructure, it's a risky bet for you. Let alone revenue recognition problems.
- Finally, these 12 attributes need to be over-communicated to your customers to properly set expectations, and obviously appear accordingly in the Terms and Conditions of a contract to avoid unwanted consequences.

While we are barely scratching the surface, the 12-attribute taxonomy illustrated by four concrete examples offers a stark illustration of both the complexity and the implied deep strategic ramifications of modern credit design.

CHAPTER 5

The C-suite implications of the 12 attributes of Credit Model

Before diving into the 4 pillars and their associated 12 attributes, it is important to reconfirm that in a Credit Model, you are purchasing a "universal currency" that is effectively a prepaid reserve of value. This universal currency is exchanged for various entitlements of “things” at different “exchange rates” (e.g., 1 credit for an image, 25 credits for a video), effectively decoupling the purchase price from the specific utility.

Figure. COMPASS' framework 12-attribute of credit model architecture (simplified version, reminder from Chapter 4).

In contrast and as seen in Chapter 3’s Flash Triage tool, in a Usage-based or Committed Model with Overage you typically meter 1 unit in most cases, and up to 3 units to keep it manageable. Customers are therefore pre and/or post-paying for a specific volume of a uniform unit where the unit itself is the product. This means that even though several of the attributes or conclusions of this article are applicable, we need to distinguish Usage and Committed Model from Credit Model per se.

Also, while we have primarily defined credits as "storage units of future utility," they take on distinct strategic identities depending on the internal stakeholder:

- For Product Teams: They are **shock absorbers** for uncertain value. In the "messy reality" of AI, compute costs are spiky and the actual ROI is often a mystery to both buyer and seller. Credits decouple the vendor's variable effort from the user's unpredictable need, helping create a much-needed "learning zone" that lets you ship powerful features without gambling on the dying per-seat model or the elusive outcome-based model for most (cf. COMPASS framework here).
- For Revenue Teams: They are **commercial friction removers**. Credits eliminate the "commitment gap" for buyers who are unable or unwilling to commit to fixed-volume contracts. They allow customers to bypass the penalty of a massive upfront "guess" and enter a discovery phase, testing both volume requirements and the utility of a wide range of capabilities without misaligned commitments.
- For CFOs: They are **contractual "quasi-currencies"**. Credits are deferred revenue liabilities and prepaid assets with volatile internal and external valuations. This requires strict exchange-rate and governance to protect margins and ensure the balance sheet remains "auditable".

Pillar I Deep-Dive: Unit Architecture (The "What")

Determines the credit definition and the fundamental rules of value exchange, i.e. what credit "buys".

1. **Valuation Modality:** The underlying anchor of consumption, i.e. the things that consume credit. Is the credit pegged to Inputs ("cost-plus"/tokens pass-thru), Activities (steps/efforts), Outputs (actions/deliverables) or even outcomes (results)?
2. **Exchange Granularity:** The conversion logic/rate cards between the credit and the action. Does 1 credit always equal 1 unit, or is it a multi-weighted system (e.g., a "Basic" task costs 1 unit while a "Premium/Reasoning" task costs 10), or a complex one with multiple rating attributes? Also, is there a consumption window (e.g., deduplication rules where re-accessing the similar executions within 24 hours costs zero credits)?
3. **Fungibility:** The breadth of the credit's purchasing power. Is the credit Universal across the ecosystem, or is it ring-fenced to specific agents, models, or product lines?

Executive considerations for Pillar I: Unit Architecture

Chief Product & Technology Officer:

- Success here hinges on **cognitive load**. If you move from a simple to a multi-weighted Granularity, the user loses the mental "price anchor" for an action. Although it requires significantly more design work, Universal Fungibility is your strongest CX lever; it removes the anxiety of "buying the wrong thing". However, the more complex your Granularity becomes, the more you risk the product feeling like a casino where the house always wins. Calibrating the number of credits used by different actions is one of the most difficult parts of the design.

Chief Revenue Officer:

- CRO needs to ensure the continuous alignment and adaptation of credit consumption to the value provided across all product lines. This requires continuous "value calibration" of the number of credits required for each service execution. Also, high Fungibility is a rep's best friend as it allows them to sell a "strategic bucket" rather than specific capabilities. It can help dramatically speed up the discovery phase. However, if the Unit Architecture is too granular or has a complex weight system, the "Ramp to Quote" time for new reps increases because they can't easily explain the value of $10k worth of credits to a prospect without a complex calculator.

Chief Financial Officer:

- From a revenue recognition perspective, this Pillar defines the Substance/Nature of the Performance Obligations (POB). Universal Fungibility (Attribute 3) generally allows for a single, broad POB, a "right to receive unspecified future goods", which simplifies initial balance sheet entry but complicates Gross Margin tracking by product line. If credits are used to buy disparate things (e.g., software vs. professional services), you face "Allocation of Transaction Price" hurdles; you must establish a verifiable Standalone Selling Price (SSP) for every possible output to ensure revenue is recognized at the correct value upon consumption. Credits act as the granular unit of measure for the Transfer of Control, turning abstract AI 'value' into a recognizable accounting event.
- Finally, it's worth noting that Pillar 1 defines the unit value at which that revenue is recorded, while as we will see, Pillar 3 "Temporal Lifecycle" defines when you can finally recognize the revenue for unused credits.

"Reality Check" for Pillar I as of March 2026

- All offers use either Activities (e.g. Tokens, Processing Units) or Outputs (e.g. Actions, Conversations) as a basis for Credits valuation modality – i.e. what "consumes" credits.
- While "synthetic credits" represent the vast majority of the market today, we are starting to see more and more "monetary credits" popping up: Anthropic or Replit are clear examples.
- This makes perfect sense given that (1) as we move from "Copilots" to "Agents," the unit shifts towards work-oriented countable units and (2) outcome-based models are usually so advanced that the granular management of credits as a proxy is almost always irrelevant, replaced by the direct negotiation of strategic KPIs, a holistic focus on total cost of ownership, and the realization of specific business transformations.
- For activity-based metrics, there has been a recent rise of synthetic metrics as an attempt to create a normalized abstraction layer of underlying pricing elements (e.g. Devin's Agent Compute Unit, ONA's Ona Compute Unit, Microsoft Secure Compute Unit, etc.). This creates additional complexity for procurement as they are not standardized nor comparable work units per se.
- The majority of vendors leverage simple rate cards / relatively limited granularity to ensure readability and forecastability. Larger players often have richer tables: Salesforce has a relatively straightforward table with approximately 12 rate cards (with several simply differentiating between production and sandbox environments), while ServiceNow describes across several pages 60+ rate cards ranging from 1 "Now Assist" for an alert analysis to 4,000 "Now Assist" for an advanced auto evaluation.
- Many companies have started providing straightforward evaluators/calculators (e.g. this one from Google) to help assess transparent credit needs as a first step.
- Although examples are limited, larger companies are progressively transitioning from Product Line fungibility to Universal fungibility by using the concept of Credit Wallet. As mentioned, ServiceNow evolved from a product-specific model where Now Assists were "ring-fenced" per product line (ITSM, HRSD...) and moved to a universal model to reduce "shelfware" and encourage cross-platform adoption. SAP followed a similar move, implementing "AI Units" that can be applied flexibly across solutions: SAP S/4HANA Cloud, SAP SuccessFactors, and SAP Ariba.

Bottomline for Pillar I:

Pillar I defines the "Value Density" of your platform. The goal is to maximize Universal Fungibility to simplify the customer's buying journey while maintaining enough Granularity to protect margins across diverse AI workloads. If the architecture is too "flat," you risk margin compression on high-cost tasks; if it is too "weighted," you create a cognitive tax that slows down sales and confuses financial auditing.

Pillar II Deep-Dive: Commercial Governance (The "How")

Ensures financial predictability for the customer and margin protection for the vendor, i.e. what credit "costs".

4. **Contractual Elasticity:** The relationship between volume commitment and unitary prices. Is the monetary-to-credit exchange rate locked for the contract duration or variable? Also, is the dollar price per credit flat, or is it a variable system (such as volume or tiering-based pricing)? If variable, is it a standard Tiered pricing for public plans and/or bespoke Negotiability for Enterprise rate cards?
5. **Economic Stability:** The protection against "Purchasing Power Risk." Is the credit-to-action exchange rate locked for the contract duration, or does the vendor reserve the right to adjust the "Rate Table"? If there is a right to adjust, is there a frequency for updates or a dynamic "Floating Rate Table" adjusting the credit cost of an action based on real-time factors like demand surge, latency, or model scarcity?
6. **Observability:** The degree of mutual transparency into consumption velocity and balance integrity. What is the frequency of utilization and credit reporting (nightly batch, real-time...), as well as proactive alerting? What is the depth of reporting, including the "math trust" of the underlying credit "rating"? What are the "wallets" for admin and user self-management? What is the data accessibility approach (APIs or dashboards) to maintain audit-ready, mutual accountability?

Executive considerations for Pillar II: Commercial Governance:

Chief Product & Technology Officer:

- High Economic Stability (Attribute 5) is essential to prevent "usage paralysis"; if the "Rate Table" is dynamic or floating, users will stop experimenting with high-value features for fear of a sudden "price surge" mid-workflow. Observability (Attribute 6) is the ultimate trust anchor: if it is "batched" or "delayed," users will perceive the credit model as a "tax" rather than a utility. Proactive alerts to both economic and business buyers are of the essence. To bridge these attributes, leveraging a robust, off-the-shelf "mediation engine" is essential. This "ledger-grade" translator converts raw, high-velocity telemetry into auditable credit events. It eliminates the revenue leakage and unscalable complexity inherent in "homegrown" systems that too often divert CTO resources while creating revenue leakage and long-term audit risks for the CFO.

Chief Revenue Officer:

- Contractual Elasticity (Attribute 4) allows Sales to defend price-per-unit for Enterprise deals, but **Bespoke Negotiability** creates massive operational debt. While tempting, multiplying bespoke rate tables will require exponentially complex setups and post-deal audits, which also slow down the entire process and, ultimately, the sales commission cycle. To scale, Sales Ops must enforce "Standard Tiers" that allow for automated billing and commission triggers based on the monetary-to-credit exchange rate, while bespoke rate tables should be limited to the largest deals when it makes a difference. Finally, while total flexibility might not be manageable, it is important to also avoid long-term price locking, as today's unit economics will likely look very different in 24 or 36 months for both sides.

Chief Financial Officer:

- This Pillar governs Revenue Recognition integrity and the management of Variable Consideration. A "Floating Rate Table" (Attribute 5) creates a significant burden for the Order-to-Cash cycle, as Finance must track the daily impact on the transaction price to assess variable consideration under ASC 606/IFRS 15. This complexity is only justified in high-volatility, tight-margin scenarios where protecting gross margin against spikes in compute cost is a survival requirement. Outside of these specific cases, the administrative overhead required to manage such fluctuations can quickly evolve from a pricing strategy into a direct operational profit drain, essentially costing more to manage than it saves in margin.
- Conversely, a "Locked Rate Table" should typically be the preferred standard; it establishes a fixed price-per-unit for the contract duration, providing a predictable revenue stream and reducing the need for constant, manual "contract modifications". Regardless of the rate structure, Auditability (Attribute 6) is a non-negotiable requirement for financial governance, made much more complex by the introduction of "credit as an intermediary currency". The system must maintain a nested "ledger-grade" audit trail where back-end consumption (Actions) is both finely measured and reconciled against the credit burn, which is in turn reconciled perfectly against the dollar value recognized on the financial statements. Any delta between the "quasi-currency" consumed and the revenue recognized introduces material risk during year-end audits.
- For the CFO, the integrity of the credit model relies on proving that every unit of utility burned correlates to a legally defensible and verifiable revenue event. Without a robust mediation system to provide pervasive proof of the Transfer of Control through telemetry, and a recurring order-to-cash platform that "natively understands" contract events, revenue recognition can quickly become a nightmare.

"Reality Check" for Pillar II as of March 2026

- For most large vendors, the monetary-to-unit rate is typically locked for the contract term.
- However, most large vendors also reserve the right to adjust the Credit-to-Utilization exchange rate (the "Rate Table"). While they are far away from "real-time dynamic"/"Uber Surge"-like rating, the rate table is effectively Floating.
- Procurement teams will therefore progressively want to include "Credit Rate Table Protection" clauses, which cap the increase in credit cost for existing OOTB skills to once per year tied to a flat percentage or CPI-like index.
- We are already witnessing some companies that have clear public guidelines on credit volume discounting. For example, Microsoft has a 5% to 20% tiered discounts structure for Copilot Studio pre-purchased credit plans.
- From an observability standpoint, close real-time reporting has become table stakes. However, only a handful of players today provide proactive credit consumption alerting or forecasting capabilities (cf. HubSpot illustration in Chapter 4, within the “Four real-life AI credit offer” subsection).
- We can safely assume that in the coming months, some companies will go beyond reporting and alerting and start also modeling the value associated of each credit, so buyers know exactly what they’re paying for and why.

Bottomline for Pillar II

Pillar II defines the "Contractual Trust" of the relationship. The goal is to maximize Stability and Observability to prevent user "usage paralysis" while strictly governing Elasticity to protect margins. While dynamic "Floating Rate Tables" offer the ultimate margin hedge, they carry a heavy Administrative Tax for you and a painful cognitive load for your customers. Unless you are operating in a hyper-volatile, low-margin environment, the "Golden Path" is a Locked Rate Table backed by a Ledger-Grade Audit Trail to ensure revenue integrity and sales velocity.

Pillar III Deep-Dive: Temporal Lifecycle (The "When")

Defines the lifecycle and lifespans of the credits.

7. **Rechargeability:** The refill mechanisms. Is it a Prepaid Draw-down (one-time purchase), a Recurring Allowance (monthly grant), or even a Threshold-based Auto-top-up (automatic refill when low)?
8. **Persistency:** The logic for unused capacity. Do credits "die" at the end of a cycle (Use-it-or-lose-it), or do they carry over to the next period? If they roll over, is there a "Cap" on the total accumulated balance? And in which order are credits consumed if they have different "vintage" / were rolled-over (e.g. "First-In, First-Out" or "Last-In, First-Out")?
9. **Terminal Expiry:** The absolute "Shelf Life" of a credit. This defines the date of final expiration from the moment of purchase, regardless of rollover status or plan activity, primarily used to manage deferred revenue liabilities on the balance sheet. This attribute also dictates the Settlement Logic: upon expiry or termination, is the value captured as high-margin "Breakage Revenue" (non-refundable) or settled via a cash-out/credit-back (partial or full Refundability)?

Executive considerations for Pillar III: Temporal Lifecycle

Chief Product & Technology Officer:

- Persistency (Attribute 8) is your primary "Fairness" metric. A strict "use-it-or-lose-it" model is viewed as a vendor-centric penalty, whereas capped rollovers foster a customer-centric partnership. Roll-over policies also lead to the complexity of defining consumption order ("First-In, First-Out" or "Last-In, First-Out"), and therefore grouped credit vintages if not individual credit lifecycle management.
- Terminal Expiry (Attribute 9) is a necessary "bad cop", typically 12 months and rarely up to 24 months, that must be communicated with extreme transparency as previously mentioned. If a user sees a balance disappear without warning, you don't just lose a credit, you lose the customer's trust in the entire ecosystem.

Chief Revenue Officer:

- Rechargeability (Attribute 7) is the engine of Net Revenue Retention (NRR). Moving from manual Draw-downs to "Automated Top-ups" shifts the Sales team from "transaction hunters" to "expansion managers". However, experience shows that forcing "Automated Top-ups" by default can alienate customers by forcing them to continuously adjust their plans. Allowing them to have on-demand top-ups, that are often 10%-25% more expensive on a unit basis as a trade-off for flexibility, can actually be a good solution to increase both retention and average deal size. This requires a radical change in commission structures: rewarding reps for "Consumption Velocity" (how fast the customer burns through credits) rather than just the initial "Big Bang" contract booking.

Chief Financial Officer:

- This Pillar defines the **Contractual Boundary** and governs the **Deferred Revenue Liability**. Unused credits stay on the balance sheet as a "Contract Liability" and can only be moved to the P&L through either what is definable as **Consumption/ Usage** (in Pillar 1) or **Breakage** (in Pillar 3). Finance teams must therefore estimate and track "Breakage Revenue", the portion of credits expected to go unused before expiry. Under ASC 606/IFRS 15, if they have sufficient historical data, they recognize this breakage proportionally as other credits are consumed. To be able to get that data, the entire Order-to-Cash chain will potentially have to manage and report individual credit lifecycles: to which period is attached this specific credit? Has it been rolled over? What is its terminal expiry?

- Without that data, they are forced to wait until the **Terminal Expiry** (Attribute 9) date is reached, which can lead to massive, unpredictable revenue "spikes" at year-end or contract-end that distort the company's actual performance to investors. In short, if you wait for the spike at the end, you are essentially "recognizing the customer's failure to use what they paid for" as a one-time gain. This is a "low-quality" revenue signal to the board and investors.

"Reality Check" for Pillar III as of March 2026

- 46% of vendors offer on-demand top-up options as one of their rechargeability mechanisms.
- Top-up credits are typically priced 20% to 50% more than subscription credits, with 30% appearing to be the sweet spot.
- Surprisingly, less than 15% of vendors offer any form of rollover today, concentrated once again with data platforms, AI APIs, as well as AI coding, and AI voice providers. This is clearly poised to increase in the coming months.
- Terminal expiry for the vast majority of vendors is either monthly or 12 months aligned with billing cycles today.
- Lovable is an interesting case: the company has recently enriched its model to include "one-off" top-ups with a 12-month rolling expiry that resets with every new purchase, creating a "quasi-evergreen" reserve of value. It is worth noting that this longevity is strictly anchored to an active subscription; while monthly and annual plans permit 1-month and full-term rollovers respectively, all accumulated credits expire at the end of the final billing cycle upon cancellation of the base plan.

Bottomline for Pillar III:

Pillar III defines the "Velocity and Liability" of the credit model. The goal is to balance Persistency (customer fairness) with Terminal Expiry (financial hygiene). While customers value "Perpetual" credits, the CFO requires a "Hard Stop" to clear liabilities, recognize Breakage Revenue, and ensure margin control. This will prevent "Zombie Liabilities". To scale, the business will ideally move from lumpy manual credit recharge to automated top-ups, shifting the revenue engine from lumpy transactions to predictable, usage-based growth.

Pillar IV Deep-Dive: Access & Entitlement (The "Who")

It specifies the organizational boundaries and permissions of the credit wallet, and the protocols for continuity when resources are exhausted.

10. Poolability: The ownership boundary. Does the credit pool live at the User level (individual), the Workspace level (Team/Department), or the Enterprise level (Shared Organization pool)?
11. Portability: The movement of ownership between distinct entities. Can credits be reassigned between different business units, legal subsidiaries? Can they even be "gifted" to third-party partners (for example as a viral trial mechanism)?
12. Overage Contingency: The "Zero-Balance" protocol. This defines the technical and commercial response when the credit bucket is exhausted, clarifying the transition from primary consumption to a contingency state. Is there a Hard Cap (service stop), a Soft Cap (throttling) or an emergency pre-defined "allowance for negative credits" / overdraft right from future periods (automatic shift to higher-priced On-Demand billing)? Or even an occasional forgiveness for the overdraft within certain limits?

Executive considerations for Pillar IV. Access & Entitlement:

Chief Product & Technology Officer:

- Poolability (Attribute 10) is the "North Star" for Enterprise UX; friction occurs when an individual is blocked while their department has an overflowing credit bucket. Overage Contingency (Attribute 12), specifically "Soft Caps", prevents the most common cause of churn: the "Service Stop". Providing a potential emergency overdraft right turns a billing crisis into a "continuity feature," making the platform feel indispensable.

Chief Revenue Officer:

- Portability (Attribute 11) can be a strategic retention lever. For example, if a rep can help a customer reassign credits from a stalled project to a high-growth business unit, they effectively save the renewal. However, "External Gifting" must be strictly governed to ensure it doesn't cannibalize the pipeline, mess up lead attribution or even create legal/compliance risks.
- Finally, to handle new client objections regarding credit volatility, calculators/simulators in pre-sales, as well as a centralized "Credit Spend Control Tower," should be table stakes. They empower customers with automated circuit breakers and spend ceilings to preemptively halt runaway agent consumption and ensure predictable governance.

Chief Financial Officer:

- This pillar creates the Entity-Level Billing Architecture. If Poolability (Attribute 10) allows for a "Global Reservoir," the legal contract must specify the "Primary Obligor" to manage tax and help anticipate intercompany transfer-pricing risks between subsidiaries.
- From an Order-to-Cash perspective, Overage Contingency (Attribute 12) must be clearly codified in the MSA; an "Emergency Overdraft" is an uncommitted purchase, and if not tied to an explicit "On-Demand" rate, it can lead to uncollectible Accounts Receivable (AR) if the customer disputes the automated invoice. Finally, occasional forgiveness introduces a "revenue leakage" risk that requires very strict policy definition, thresholds and automated tracking to ensure these unbilled consumption events are treated as strategic marketing investments (contra-revenue) rather than an unmanaged erosion of gross margin.

"Reality Check" for Pillar IV as of March 2026

- 60% of offers do not offer pooling at all, i.e. credit stays at user level. When the offer does offer pooling, it's worth noting that it is concentrated within the segments of platforms that are not user-based by design (API platforms, Data Platforms...). Simply put, credit pooling though growing remains a minority. This comes as a surprise as pooling typically helps ensure "quality revenue"
- The majority of large vendors have transitioned from rigid, siloed entitlements to "Enterprise Digital Wallets" at the Global Account level, allowing for seamless internal reassignment across all business units and legal subsidiaries under a single global agreement.
- There are emerging cases of cross-company pools and transfers. ServiceNow allows MSPs to manage a "Universal Pool" of Assists across multiple of their customer instances. In some 2026 contracts, the MSP might be able to "pass through" a starter bucket of their own Assists to a new client to prove the ROI of a specific AI skill (like Summarization) before the client commits to their own Assist Pack.
- Regarding overage, almost 50% of offers have a hard stop. While "soft caps with overage" are gaining traction, frictionless top-up mechanisms should also help improve UX on this point.
- ServiceNow has introduced an AI Control Tower that helps enforce governance through departmental "circuit breakers" that halt execution once spend ceilings are reached. Complementing this, Now Assist Guardian utilizes rate-limiting to provide a technical safety net against "runaway agents" and rapid credit depletion.

Bottomline for Pillar IV:

Pillar IV defines the "Operational Friction" of the ecosystem. The goal is to maximize Poolability to ensure credits flow to where value is being created, while strictly defining Overage Contingencies to prevent service interruptions. For the organization, this pillar represents a shift from managing "Individual Users" to managing "Entitlement Reservoirs", requiring a legal and technical framework that supports fluid usage without creating "Uncollectible Revenue" or tax-compliance risks.

CHAPTER 6

Migrating towards a new Credit-based Model: Not a walk in the park

Moving to a credit-based model is a high-stakes structural transformation, not a simple pricing update. As the COMPASS credit-based model 12-attribute framework suggests, this transition triggers deep ramifications across the enterprise. To name a few, the CPTO must re-architect the product for granular telemetry, while the CRO must equip teams with updated playbooks to navigate new commercial frictions and "commitment anxiety."

However, the CFO faces by far the greatest struggle; credits transform predictable recurring payments into complex deferred revenue liabilities, creating an exponential "Order-to-Cash" nightmare for teams without ledger-grade infrastructure to reconcile usage with revenue recognition. This operational tension is further exacerbated by a looming human resource crisis: more than 7 out of 10 CPAs are nearing retirement, and the U.S. has lost more than 300,000 accountants these past years.

In this environment, CFOs simply cannot solve the credit complexity problem by "throwing more bodies" at it. Instead, they must urgently prioritize their order-to-cash and revenue management platform readiness, including AI capabilities that have a zero-error threshold. This will help ensure that the required agility of their monetization strategy does not outpace their ability to maintain financial integrity.

Beyond these internal shifts and system upgrades, the external customer experience is where many companies learn, often the hard way, that transparency is an invaluable lever. High-performing migrations prioritize customer-centric "predictability tools" to bridge the conceptual gap before the transition, typically including:

- **Credit forecasting tools** to help users assess their future needs, for both new and existing customers.
- **Credit utilization simulators** to bridge the logic between legacy units and new credits, for existing customers.
- **Credit spend controls** so customers can define overage policies before they trigger.
- **Proactive threshold alerts** to prevent "bill shock."
- **"Shadow Billing" period(s)**, before the new credit model goes live ("Under your current plan, you paid $X. Under the upcoming credit model, you would have consumed Y credits").
- Ultimately, these tools are for both the buyers and sales teams. They turn the "Credit Model Blending and Calibration Rules of Thumb" previously mentioned into an actionable reality, ensuring rightsized allotments for both new market acquisitions and legacy customer migrations.

Cursor provides a striking case in point about "learning the hard way". In June 2025, the company materially transformed its monetization model, moving from a limit of "500 fast requests per month" to a "$20 worth of credits value" bucket (i.e. monetary credit). The challenge was that Cursor did not over-communicate this change or provide significant notice. Crucially, they did not proactively provide easy ways for users to estimate how their specific usage would consume credits in the new model. This was compounded by potentially conflicting messaging across blog posts and pricing pages.

As my friend and expert Steven Forth simply puts it:

> “
>
> Predictability and Transparency trump precision.
>
> **Steven Forth**
> Founder, valueIQ.ai

Many users discovered the change only after hitting unexpected usage limits. In fact, heavy users of premium models like Claude Opus ran out of credits within days, leading to surprise overages and widespread frustration on social media. To his credit, CEO Michael Truell reacted swiftly and with high integrity; he issued a public apology and offered immediate refunds, acknowledging, "We didn't handle this pricing rollout well." While his leadership saved the community's trust, the story remains a powerful cautionary tale. It reminds us of the absolute importance of treating a credit migration not just as a math problem, but as a transparency exercise.

Strategically, as AI credits will command an ever-larger slice of enterprise budgets, buyers will likely strengthen their FinOps muscle. Vendors will be compelled to abandon the monetization “black box” for a “glass-box” model: delivering real-time observability via standards like MCP, event-based alert triggers and hyper-granular audit trails rather than basic dashboards. This mandate could become binary: if a customer cannot directly access your usage events to drive their own internal modeling and forecasting, the contract will not survive the renewal cycle. This internal demand mirrors a new external reality: visibility is shifting from human-led search to AI discovery. In an agentic economy, if your unit economics aren't machine-accessible and defensible at the granular level, your brand is effectively invisible to the procurement agents that will sooner or later control the purse strings. While this is true for any software, credit-based models are naturally harder to track, making the transparency mandate even more critical.

Is it time to design your own credit models and which questions should you ask?

The return of the credit model marks a pivotal transition in this brave new world of AI. While credits have deep roots in the "old guard" of telecommunications and information providers, they have been resurrected as a monetization stabilizer for the agentic era. While we don't have a crystal ball to predict whether credits are the destination, we can confidently view them at the very least as a Strategic Bridge that will remain essential for the foreseeable future. They help buyers and sellers progressively tame what "value" actually looks like in an agentic world. And big names such as OpenAI, Anthropic, and HubSpot are already doing the work of market education.

As expert and friend Mark Stiving, Ph.D. points out:

> "Credits sit above tokens. They make pricing understandable and approachable for buyers who don't think in tokens or inference calls.
>
> **Mark Stiving, Ph.D.**
> Author of *Impact Pricing*

However, leaders must move with caution. Credits can easily become an unnecessary "layer of abstraction," introducing commercial friction and shaving margins through operational drag. To navigate this, your design process must be rooted in architectural rigor, not just competitive imitation. Before committing to this path, executives should start by leveraging this article's **COMPASS Credit Flash Diagnostic** to ensure their context truly demands a credit substrate. If confirmed, then this article's **COMPASS Credit Model 12-Attribute Taxonomy** provides the necessary scaffolding to avoid the "zombie liabilities" and "usage paralysis" design blind spots that plague poorly designed offers.

The most successful architectures we see today are not "all-or-nothing" bets; we are already in an era of Monetization Hybridization. By layering credits onto pre-existing seat-based models or platform fees, organizations can hedge their revenue baskets: they can capture the "spiky" upside of AI utility while maintaining the recurring predictability that investors demand. Moreover, hybridization is inherently linked to increased complexity, requiring a structured, systematic framework. The COMPASS framework includes tools such as the **Blending and Calibration Rules of Thumb** (introduced in this article), alongside established ones such as the **COMPASS 3x3 Metric Matrix** and the **COMPASS 7-Fold Star**, to stress-test your models against the framework's 14 specific "Fit Check" criteria from both the vendor and buyer perspectives.

Ultimately, a shift to credits is not just a change on the price list; it is a structural transformation of the enterprise. The shift to credit-based monetization demands a new, interlocking alignment: the CPTO must lead the evolution of the product experience, the CRO must navigate new commercial frictions, and the CFO must safeguard financial integrity against exponential complexity.

Success requires, more than ever, a shift in perspective from departmental silos to a unified Quote-to-Cash strategy. Without a ledger-grade infrastructure to reconcile telemetry with revenue recognition, the 'dream' of credit flexibility will quickly devolve into a financial nightmare of reconciliation discrepancies and operational risk. To address the "AI credit" paradigm, your finance platform readiness must finally match your revenue architecture ambition.

In this era of unprecedented iteration, agility is your greatest asset. There is no such thing as a "perfect" or "set-and-forget" monetization strategy; I have seen more iterations and changes in the last three years than in the previous twenty, and the pace is only poised to accelerate. Stay open, stay agile, and remember: while your credit-model design can help protect your margins, your customers are the ultimate judges of your value. Build a bridge that is as resilient as it is flexible.

Zuora provides the leading monetization platform and system of record for quote-to-cash, helping companies operationalize complex revenue models, including usage, hybrid, and AI-driven pricing, without breaking finance. As AI changes what companies sell, how customers buy, and how revenue is generated and managed, Zuora's flexible, modular software solutions help businesses adapt pricing and packaging, billing, payments, and revenue recognition to support increasing scale and complexity. Customers around the world, including BMC Software, Box, General Motors, The New York Times, Schneider Electric, and Zoom use Zuora's unique combination of technology and expertise to transform their financial operations and how they go to market. Zuora is headquartered in Silicon Valley with offices in the Americas, EMEA, and APAC. To learn more, please visit zuora.com.

ZUORA